ACCELERATED AND INTERPRETABLE OBLIQUE RANDOM SURVIVAL FORESTS

A PREPRINT

Byron C. Jaeger
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA
bjaeger@wakehealth.edu

Sawyer Welden
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA
swelden@wakehealth.edu

Kristin Lenoir
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA
klenoir@wakehealth.edu

Jaime L. Speiser
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA
jspeiser@wakehealth.edu

Matthew Segar
Department of Cardiology
Texas Heart Institute
Houston, TX 77030, USA
Matthew.Segar@BCM.edu

Ambarish Pandey
Division of Cardiology, Department of Internal Medicine
University of Texas Southwestern Medical Center
Dallas, TX 75235, USA
Ambarish.Pandey@UTSouthwestern.edu

Nicholas M. Pajewski
Department of Biostatistics and Data Science
Wake Forest University School of Medicine
Winston-Salem, NC 27157, USA
npajewsk@wakehealth.edu

July 27, 2022

ABSTRACT

The oblique random survival forest (RSF) is an ensemble supervised learning method for right-censored outcomes. Trees in the oblique RSF are grown using linear combinations of predictors to create branches, whereas in the standard RSF, a single predictor is used. Oblique RSF ensembles often have higher prediction accuracy than standard RSF ensembles. However, assessing all possible linear combinations of predictors induces significant computational overhead that limits applications to large-scale data sets. In addition, few methods have been developed for interpretation of oblique RSF ensembles, and they remain more difficult to interpret compared to their axis-based counterparts. In this article, we introduce and evaluate a method to increase computational efficiency of the oblique RSF and a method to estimate importance of individual predictor variables with the oblique RSF. Our strategy to reduce computational overhead makes use of Newton-Raphson scoring, a classical optimization technique that we apply to the Cox partial likelihood function within each non-leaf node of decision trees. We estimate the importance of individual predictors for the oblique RSF by negating each coefficient used for the given predictor in linear combinations, and then computing the reduction in out-of-bag accuracy. In general benchmarking experiments, we find that our implementation of the oblique RSF is over 500 times faster with equivalent discrimination and superior Brier score compared to existing software for oblique RSFs. We find in simulation studies that 'negation importance' discriminates between signal and noise predictors more reliably than

permutation importance, Shapley additive explanations, and a previously introduced technique to measure variable importance with oblique RSFs based on analysis of variance. All methods pertaining to oblique RSFs in the current study are available in the `aorsf` R package.

Keywords Random Forests · Survival · Efficiency · Variable Importance

1 Introduction

Risk prediction may reduce the burden of disease by guiding strategies for prevention and treatment in a wide range of domains [Moons et al., 2012a,b]. The random survival forest (RSF; Ishwaran et al. [2008], Hothorn et al. [2006]) is a supervised learning algorithm that has been used frequently for risk prediction [Wang and Li, 2017]. Similar to random forests (RFs) for classification and regression [Breiman, 2001], The RSF is a large set of de-correlated and randomized decision trees, with each tree contributing to the ensemble's prediction function. Notable characteristics of the RSF include uniform convergence of its ensemble survival prediction function to the true survival function, first shown by Ishwaran and Kogalur [2010] and later by Cui et al. [2017] under more general conditions. However, Cui et al. [2017] noted that the RSF is at a disadvantage when predictors are correlated and some are not relevant to the censored outcome, which is a strong possibility when large medical databases are leveraged for risk prediction.

A potential approach to improve the RSF when predictors are correlated and some are not relevant to the censored outcome is to use oblique trees instead of axis based trees. Axis based trees split data using a single predictor, creating decision boundaries that are perpendicular or parallel to axes of the predictor space [see Breiman et al., 2017, Chapter 2]. Oblique trees split data using a linear combination of predictors, creating decision boundaries that are neither parallel nor perpendicular to axes of their contributing predictors [see Breiman et al., 2017, Chapter 5]. Menze et al. [2011] examined prediction accuracy of RFs in the presence of correlated predictors and found that oblique RFs had substantially higher prediction accuracy compared to axis-based RFs. Similarly, Jaeger et al. [2019] found that growing RSFs with oblique rather than axis-based survival trees reduced the RSF's concordance error, with improvements ranging from 2.5% to 24.9% depending on the data analyzed.

Oblique trees have at least two notable drawbacks compared to axis-based trees. First, finding a locally optimal oblique decision rule may require exponentially more computation than an axis-based rule. If p predictors are potentially used to split n observations, up to $\mathcal{O}(n^p)$ oblique splits can be assessed versus $\mathcal{O}(n \cdot p)$ axis-based splits [Heath et al., 1993, Murthy et al., 1994]. Second, estimating variable importance (VI) using permutation (a standard method for RFs) may be less effective in ensembles of oblique trees, as permuting the values of one predictor may not destabilize decisions that are based on linear combinations of predictors. Although VI is one of the most widely used strategies to interpret random forests [Ishwaran and Lu, 2019], few studies have investigated VI for oblique random forests [see Menze et al., 2011, Section 5], and fewer have investigated VI specifically for the oblique RSF.

The rest of this article is organized as follows. Section 2 reviews prior studies that have developed methods related to those introduced in the current study. In Section 3, we reduce the computational cost of oblique RSFs (that is, accelerate them) with a scalable algorithm to identify linear combinations of coefficients. In Section 4, we improve the interpretability of oblique RSFs with 'negation VI', a method to estimate VI that flips the sign of coefficients in linear combinations of predictors instead of permuting predictor values. We evaluate these methods with general benchmarking experiments and simulation studies in Section 5. In Section 6, we summarize results from the current study and present ideas connecting the current work to existing frameworks and methods for RSFs that future studies may engage with. The accelerated oblique RSF and multiple methods to compute VI for oblique RSFs are available in the `aorsf` R Package.

2 Related work

Sections 2.1 and 2.2 briefly summarize prior studies that have developed methods related to the oblique RSF and VI, respectively.

2.1 Axis-based and oblique random forests

After Breiman [2001] introduced the axis-based and oblique RF, numerous methods were developed to grow oblique RFs for classification or regression tasks [Menze et al., 2011, Zhang and Suganthan, 2014, Rainforth and Wood, 2015, Zhu et al., 2015, Poona et al., 2016, Qiu et al., 2017, Tomita et al., 2020, Katuwal et al., 2020]. However, oblique splitting approaches for classification or regression may not generalize to censored outcomes [for example, see Zhu, 2013, Section 4.5.1], and most research involving the RSF has focused on forests with axis-based trees [Wang and Li, 2017].

Building on prior research for bagging survival trees [Hothorn et al., 2004], Hothorn et al. [2006] developed an axis-based RSF in their framework for unbiased recursive partitioning, more commonly referred to as the conditional inference forest (CIF). Zhou et al. [2016] developed a rotation forest based on the CIF and Wang and Zhou [2017] developed a method for extending the predictor space of the CIF. Ishwaran et al. [2008] developed an axis-based RSF with strict adherence to the rules for growing trees proposed in Breiman [2001]. Jaeger et al. [2019] developed the oblique RSF following the bootstrapping approach described in Breiman's original RF and incorporating early stopping rules from the CIF.

Fast algorithms to fit axis-based RSFs are available in the `randomForestSRC` R package [Ishwaran and Kogalur, 2019] and the `ranger` [Wright and Ziegler, 2017] R package. `randomForestSRC` provides a unified interface to grow RFs in a wide range of analyses, and `ranger` is designed to grow RFs efficiently using high dimensional data. Fast algorithms to fit the CIF are provided by the `party` R package [Hothorn et al., 2010], which provides a computational toolbox for recursive partitioning using conditional inference trees. Jaeger et al. [2019] developed the `obliqueRSF` package and found it was approximately 30 times slower than `party` and nearly 200 times slower than `randomForestSRC`. Few studies have developed software with fast algorithms for oblique RSFs that have comparable speed compared to algorithms for axis-based RSFs.

2.2 Variable importance

Several techniques to estimate VI have been developed since Breiman [2001] introduced permutation VI, which is defined for each predictor as the difference in a RF's estimated prediction error before versus after the predictor's values are randomly permuted. Strobl et al. [2007] identified bias in permutation VI driven by variable selection bias and effects induced by bootstrap sampling, and proposed an unbiased permutation VI measure based on unbiased recursive partitioning [Hothorn et al., 2006]. Menze et al. [2011] introduced an approach to estimate VI for oblique RFs that computes an analysis of variance (ANOVA) table in non-leaf nodes to obtain p-values for each predictor contributing to the node. The ANOVA VI[1] is then defined for each predictor as the number of times a p-value associated with the predictor is ≤ 0.01 while growing a forest. Lundberg and Lee [2017] introduced a method to estimate VI using SHapley Additive exPlanation (SHAP) values, which estimate the contribution of a predictor to a model's prediction for a given observation. SHAP VI is computed for each predictor by taking the mean absolute value of SHAP values for that predictor across all observations in a given set. With the exception of Menze et al. [2011], few studies have evaluated estimation of VI using oblique RFs, and fewer have examined VI specifically for the oblique RSF.

3 The accelerated oblique random survival forest

This section describes our approach to reduce computational overhead of the oblique RSF. Consider the usual framework for right-censored time-to-event outcomes with training data

$$\mathcal{D}_{\text{train}} = \{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N_{\text{train}}}.$$

Here, T_i is the event time if $\delta_i = 1$ or the censoring time if $\delta_i = 0$, and $\boldsymbol{x}_i$ is a vector of predictors values. Assuming there are no ties, let $t_1 < \ldots < t_m$ denote the m unique event times in $\mathcal{D}_{\text{train}}$.

To accelerate the oblique RSF, we propose to identify linear combinations of predictor variables in non-leaf nodes by applying Newton Raphson scoring to the partial likelihood function of the Cox regression model:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}}, \tag{1}$$

where R_i is the set of indices, j, with $T_j \geq t_i$ (i.e., those still at risk at time t_i), and $j(i)$ is the index of the observation for which an event occurred at time t_i. Newton Raphson scoring is an exceptionally fast estimation procedure, and the `survival` package includes documentation that outlines how to efficiently program it [Therneau, 2022a]. Briefly, a vector of estimated regression coefficients, $\hat{\beta}$, is updated in each step of the procedure based on its first derivative, $U(\hat{\beta})$, and second derivative, $H(\hat{\beta})$:

$$\hat{\beta}^{k+1} = \hat{\beta}^k + U(\hat{\beta} = \hat{\beta}^k)\, H^{-1}(\hat{\beta} = \hat{\beta}^k)$$

For statistical inference, it is recommended to continue updating $\hat{\beta}$ by completing additional iterations of Newton Raphson scoring until a convergence threshold is met. However, since an estimate of $\hat{\beta}$ is created by the first iteration of

[1] Menze et al. [2011] name their method 'oblique RF VI', but we use the name 'ANOVA VI' in this article to avoid confusing Menze's approach with other approaches to estimate VI for oblique RFs.

Newton Raphson scoring, only one iteration of Newton Raphson scoring is needed to identify a valid linear combination of predictors. Moreover, computing U and H requires computation and exponentiation of the vector $\boldsymbol{x}\hat{\beta}$, but these steps can be skipped on the first iteration of Newton Raphson scoring if an initial value of $\hat{\beta} = 0$ is chosen, allowing for a reduction in computing operations and removing the need to scale predictor values prior to initiating the Newton Raphson algorithm.[2] In Section 5.1.6, we formally test whether growing oblique survival trees using one iteration of Newton Raphson scoring provides equivalent prediction accuracy compared to trees where iterations are completed until a convergence threshold is met.

Algorithm 1 presents our approach to fitting an oblique survival tree in the accelerated oblique RSF using default values from the `aorsf` R package. Several steps are taken to reduce computational overhead. First, memory is conserved by conducting bootstrap resampling via randomly generated bootstrap weights rather than making a traditional bootstrap sample. Weights are integer valued, with a weight of v indicating an observation was sampled v times. Second, early stopping is applied to the tree-growing procedure if a statistical criterion is not met. In our case, the criterion is based on the magnitude of a log-rank test statistic corresponding to splitting the data at a current node. Third, instead of greedy recursive partitioning, we use 'good enough' partitioning. More specifically, instead of computing a log-rank test statistic for several different linear combinations of variables and proceeding with the highest scoring option, we identify an optimal cut-point for one linear combination of variables and assess whether using this combination will create a split that passes the criterion for splitting a node. If it does not pass the criterion, then another linear combination will be tested, with the maximum number of attempts set by the parameter `n_retry`. Often a 'good-enough' split can be found in just one attempt when the training set is large, which gives the accelerated oblique RSF a computational advantage in larger training sets compared to greedy partitioning.

4 Negation variable importance

This Section introduces negation VI, which is similar to permutation VI in that it measures how much a model's prediction error increases when a variable's role in the model is de-stabilized. Specifically, negation VI measures the increase in an oblique RF's prediction error after flipping the sign of all coefficients linked to a variable (that is, negating them). As the magnitude of a coefficient increases, so does the probability that negating it will change the oblique RF's predictions. For the current study, we use Harrell's concordance (C)-statistic [Harrell et al., 1982] to measure change in prediction error when computing negation VI.

Negation VI has several helpful characteristics. First, negation VI generalizes to any oblique RF (that is, not just RSFs) using any valid error function, making it both general and flexible.[3] Second, since the coefficients in each non-leaf node of an oblique tree are adjusted for the accompanying predictors, negation VI may provide better estimation of VI in the presence of correlated variables compared to standard VI techniques. Third, unlike permutation, negation is non-random and hence reproducible without setting a random seed. Additionally, since negation VI does not permute variables, the analyst need not worry about impossible combinations of predictors that may occur when one predictor is randomly permuted, such as having a negative status for type 2 diabetes and having Hemoglobin A1c level $\geq 6.5\%$ (a value indicative of type 2 diabetes) as a result of randomly permuting the values of Hemoglobin A1c.

5 Numeric experiments

Sections 5.1 and 5.2 present numerical experiments examining the accelerated oblique RSF and negation VI, respectively. The code used to run these experiments is available online at https://github.com/bcjaeger/aorsf-bench. All analyses were conducted using R version 4.1.3 and coordinated by the `targets` R package [Landau, 2021]. To standardize comparisons of computational efficiency, all learners and VI techniques used up to 4 processing units.

5.1 Benchmark of prediction accuracy and computational efficiency

The aim of this numeric experiment is to evaluate and compare the accelerated oblique RSF with its predecessor (the oblique RSF from the `obliqueRSF` R package) and with other machine learning algorithms for risk prediction. Inferences drawn from this experiment include equivalence and inferiority tests based on Bayesian linear mixed models.

[2] Predictors are scaled prior to initiating the Newton Raphson algorithm to avoid exponentiation of large numbers. However, if only one iteration is completed with an initial value of 0 for $\hat{\beta}$, then $\exp(\boldsymbol{x}\hat{\beta}) = 1$.

[3] The `aorsf` package enables customized functions to be applied in lieu of the default C-statistic (see `?aorsf::orsf_vi_negate`)

Algorithm 1 Accelerated oblique random survival tree using default parameters.

Require: Training data $\mathcal{D}_{\text{train}} = \{(T_i, \delta_i, \boldsymbol{x}_i)\}_{i=1}^{N_{\text{train}}}$, mtry $= \sqrt{\text{ncol}(\boldsymbol{x}_{\text{train}})}$, n_split $= 5$, n_retry $= 3$, and split_min_stat $= 3.841459$

$\mathcal{T} \leftarrow \emptyset$
$w \leftarrow$ sample(from $= \{0, \ldots, 10\}$, size $=$ nrow($\boldsymbol{x}_{\text{train}}$), replace $=$ T)
$\mathcal{D}_{\text{in-bag}} \leftarrow$ subset($\mathcal{D}_{\text{train}}$, rows $=$ which($w > 0$))
$w \leftarrow$ subset(w, which($w > 0$))
node_assignments $\leftarrow$ rep(1, times $=$ nrow($\boldsymbol{x}_{\text{in-bag}}$))
nodes_to_split $\leftarrow \{1\}$
while nodes_to_split $\neq \emptyset$ **do**
    for node $\in$ nodes_to_split **do**
        n_try $\leftarrow 1$
        node_rows $\leftarrow$ which(node_assignments $\equiv$ node)
        node_cols $\leftarrow$ sample(from $= \{1, \ldots, \text{ncol}(\boldsymbol{x})\}$, size $=$ mtry, replace $=$ F)
        $\mathcal{D}_{\text{node}} \leftarrow$ subset($\mathcal{D}_{\text{in-bag}}$, rows $=$ node_rows, columns $=$ node_cols)
        $\beta \leftarrow$ newt_raph($\mathcal{D}_{\text{node}}$, weights $=$ subset(w, node_rows), max_iter $= 1$)
        $\eta \leftarrow \boldsymbol{x}_{\text{node}} \times \beta$
        $\mathcal{C} \leftarrow$ sample(from $=$ unique(η), size $=$ n_split, replace $=$ F)
        $c \leftarrow \text{argmax}_{c^* \in \mathcal{C}} \{\text{log\_rank\_stat}(\eta, c^*)\}$
        if log_rank_stat(η, c) $\geq$ split_min_stat **then**
            $\mathcal{T} \leftarrow$ add_node($\mathcal{T}$, name $=$ node, beta $= \beta$, cutpoint $= c$)
            ▷ Right node logic omitted for brevity (identical to left node logic)
            node_left_name $\leftarrow$ max(node_assignments) $+ 1$
            node_left_rows $\leftarrow$ subset(node_rows, which($\eta \leq c$))
            subset(node_assignments, node_left_rows) $\leftarrow$ node_left_name
            if is_splittable(subset(node_assignments, node_left_rows)) **then**
                nodes_to_split $\leftarrow$ nodes_to_split $\cup$ node_left_name
            else
                $\mathcal{T} \leftarrow$ add_leaf($\mathcal{T}$, data $=$ subset($\mathcal{D}_{\text{node}}$, rows $=$ node_left_rows))
            end if
        else if n_try $\leq$ n_retry **then**
            n_try $\leftarrow$ n_try $+ 1$
            go to 11
        else
            $\mathcal{T} \leftarrow$ add_leaf($\mathcal{T}$, data $= \mathcal{D}_{\text{node}}$)
        end if
        nodes_to_split $\leftarrow$ nodes_to_split $\setminus$ {node}
    end for
end while
return $\mathcal{T}$

5.1.1 Learners

We consider four classes of learners: RSFs (both axis-based and oblique), boosting ensembles, regression models, and neural networks. Specific learners from each class are summarized in Table 1. To facilitate fair comparisons, tuning parameters were harmonized within each class. For example, for RSF learners, we set the minimum node size (a parameter shared by all RSF learners) as 10. Additionally, for RSF learners, the number of randomly selected predictors was the square root of the total number of predictors rounded to the nearest integer, and the number of trees in the ensemble was 500. For boosting, regression, and neural network learners, nested 10-fold cross-validation was applied to tune relevant model parameters. Specifically, tuning for boosting models included identifying the number of steps to complete. For regression models, tuning was used to identify the magnitude of penalization. For neural networks, the number and density of layers was tuned.

Learner Class	Software	Learners	Description
Random Survival Forests			

Axis based	`RandomForestSRC`
	`ranger`
	`party`
	`rotsf`
	`rsfse`
		`rsf-standard`
		`rsf-extratrees`
		`cif-standard`
		`cif-rotate`
		`cif-spacextend`
			`rsf-standard` grows survival trees following Leo Breiman's original random forest algorithm with variables and cut-points selected to maximize a log-rank statistic. `rsf-extratrees` grows survival trees with randomly selected features and cut-points. `cif-standard` uses the framework of conditional inference to grow survival trees. `cif-rotate` extends `cif-standard` by applying principal component analysis to random subsets of data prior to growing each survival tree. `cif-spacextend` derives new predictors for each tree in the ensemble, separately.

Oblique	`obliqueRSF`
	`aorsf`
		`obliqueRSF-net`
		`aorsf-net`
		`aorsf-fast`
		`aorsf-cph`
		`aorsf-extratrees`
			Oblique survival trees following Leo Breiman's random forest algorithm. Linear combinations of inputs are derived using `glmnet` in `obliqueRSF-net` and `aorsf-net`, using Newton Raphson scoring for the Cox partial likelihood function in `aorsf-fast` (1 iteration of scoring) and `aorsf-cph` (up to 20 iterations), and chosen randomly from a uniform distribution in `aorsf-extratrees`. Cut-points are selected from 5 randomly selected candidates to maximize a log-rank statistic.

Boosting ensembles			
Trees	`xgboost`	`xgboost-cox` `xgboost-aft`	`xgboost-cox` maximizes the Cox partial likelihood function, whereas `xgboost-aft` maximizes the accelerated failure time likelihood function. Nested cross validation (5 folds) is applied to tune the number of trees grown, the minimum number of observations in a leaf node was 10, the maximum depth of trees was 6, and $\sqrt{p}$ variables were considered randomly for each tree split, where p is the total number of predictors.
Regression models			
Cox Net	`glmnet`	`glmnet-cox`	The Cox proportional hazards model is fit using an elastic net penalty. Nested cross validation (5 folds) is applied to tune penalty terms.
Neural networks			
Cox Time	`survivalmodels`	`nn-cox`	A neural network based on the proportional hazards model with time-varying effects. Nested cross-validation was applied to select the number of layers (from 1 to 8), the number of nodes in each layer (from $\sqrt{p}/2$ to $\sqrt{p}$), and the number of epochs to complete (up to 500). A drop-out rate of 10% was applied during training.

Table 1: Learning algorithms assessed in numeric studies. `aorsf-fast` is the accelerated oblique random survival forest (see Algorithm 1), and each of the additional learners are compared to `aorsf-fast` in numeric studies.

5.1.2 Evaluation of prediction accuracy

Our primary metric for evaluating the accuracy of predicted risk is the integrated and scaled Brier score [Graf et al., 1999], a proper scoring rule that combines discrimination and calibration in one value and improves interpretability by adjusting for a benchmark model [Kattan and Gerds, 2018]. Consider a testing data set:

$$\mathcal{D}_{\text{test}} = \{(T_i, \delta_i, x_i)\}_{i=1}^{N_{\text{test}}}.$$

Let $\widehat{S}(t \mid x_i)$ be the predicted probability of survival up to a given prediction time of $t > 0$. For observation i in $\mathcal{D}_{\text{test}}$, let $\widehat{S}(t \mid \boldsymbol{x}_i)$ be the predicted probability of survival up to a given prediction time of $t > 0$. Define

$$\widehat{\text{BS}}(t) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \{\widehat{S}(t \mid \boldsymbol{x}_i)^2 \cdot I(T_i \leq t, \delta_i = 1) \cdot \widehat{G}(T_i)^{-1} + [1 - \widehat{S}(t \mid \boldsymbol{x}_i)]^2 \cdot I(T_i > t) \cdot \widehat{G}(t)^{-1}\}$$

where $\widehat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution. As $\widehat{\text{BS}}(t)$ is time dependent, integration over time provides a summary measure of performance over a range of plausible prediction times. The integrated $\widehat{\text{BS}}(t)$ is defined as

$$\widehat{\mathcal{BS}}(t_1, t_2) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \widehat{\text{BS}}(t) dt. \tag{2}$$

In our results, t_1 and t_2 are the 25th and 75th percentile of event times, respectively. $\widehat{\mathcal{BS}}(t_1, t_2)$, a sum of squared prediction errors, can be scaled to produce a measure of explained residual variation (that is, an R^2 statistic) by computing

$$R^2 = 1 - \frac{\widehat{\mathcal{BS}}(t_1, t_2)}{\widehat{\mathcal{BS}}_0(t_1, t_2)} \tag{3}$$

where $\widehat{\mathcal{BS}}_0(t_1, t_2)$ is the integrated Brier score when a Kaplan-Meier estimate for survival based on the training data is used as the survival prediction function $\widehat{S}(t)$. We refer to this R^2 statistic as the index of prediction accuracy (IPA) [Kattan and Gerds, 2018].

Our secondary metric for evaluating predicted risk is the time-dependent concordance (C)-statistic. We compute the first time-dependent C-statistic proposed by Blanche et al. [2013, Equation 3], which is interpreted as the probability that a risk prediction model will assign higher risk to a case (that is, an observation with $T \leq t$ and $\delta = 1$) versus a non-case (that is, an observation with $T > t$). Similar to the IPA, observations with $T \leq t$ and $\delta = 0$ only contribute to inverse probability of censoring weights for the time-dependent C-statistic.

Both the IPA and time-dependent C-statistic generally take values between 0 and 1. To avoid presenting an excessive amount of leading zeroes in our tables, figures, and text, we scale both the IPA and time-dependent C-statistic by 100. For example, we present a value of 25 if the IPA is 0.25, 87 if the time-dependent C-statistic is 0.87, and present 10.2 if the difference between two IPA values is 0.102

5.1.3 Data sets

We used a collection of 20 data sets containing a total of 33 risk prediction tasks (tasks per data set ranged from one to four) to benchmark the accelerated oblique RSF versus other learners. Participant-level data from the GUIDE-IT and SPRINT clinical trials and the ARIC and MESA community cohort studies was obtained from the National Institute of Health Biologic Specimen and Data Repository Coordinating Center (BioLINCC). Designs and protocols for these studies have been made available [ARIC Investigators, 1989, Bild et al., 2002, Felker et al., 2017, SPRINT Research Group, 2015]. All other datasets were publicly

available and obtained through R packages (see Appendix A.1). Across all prediction tasks, the number of observations ranged from 137 to 17,549 (median: 1,384), the number of predictors ranged from 7 to 1,692 (median: 24), and the percentage of censored observations ranged from 5.26 to 97.7 (median: 72.5) (Table A.1).

5.1.4 Monte-Carlo cross validation

For each risk prediction task, we completed 25 runs of Monte-Carlo cross validation. In each run, we used a random sample containing 50% of the available data for training and the remaining 50% for testing each of the learners described in Section 5.1.1. Then, for each learner, we computed the IPA, time-dependent C-statistic, and computational time required to fit a prediction model and compute risk predictions. If any learner failed to obtain predictions on any particular split of data[4], the results for that split were omitted from downstream analyses.

5.1.5 Statistical analysis

After collecting data from 25 replications of Monte-Carlo cross validation for the 14 learners in all 33 risk prediction tasks, we analyzed the resulting 11,550 observations of IPA and, separately, time-dependent C-statistic, using a Bayesian linear mixed model. Our approach follows the ideas described by Benavoli et al. [2017] and Kuhn and Wickham [2020], who developed guidelines on making statistical comparisons between learners using Bayesian models. Specifically, we fit two models:

$$\text{IPA} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run})$$

and

$$\text{C-stat} = \widehat{\gamma}_0 + \widehat{\gamma} \cdot \text{learner} + (1 \,|\, \text{data/run}).$$

Random intercepts for specific splits of data (that is, `run` in the model formula) were nested within datasets. The intercept, $\widehat{\gamma}_0$, was the expected value of the outcome using `aorsf-fast`, making the coefficients in $\widehat{\gamma}$ the expected differences between `aorsf-fast` and other learners. Default priors from `rstanarm` were applied for model fitting [Goodrich et al., 2022].

Hypothesis testing For both the IPA and time-dependent C-statistic, we conducted equivalence and inferiority tests based on a 1 point region of practical equivalence. More specifically, we concluded that two learners had practically equivalent IPA or time-dependent C-statistic if there was a 95% or higher posterior probability that the absolute difference in the relevant metric was less than 1. We concluded that one learner was weakly superior when there was $\geq$ 95% posterior probability that the difference in the relevant metric was non-zero, and concluded superiority when when there was $\geq$ 95% posterior probability that the difference in the relevant metric was 1 or more.

5.1.6 Results

A full summary of all results presented in this Section is provided in Table A.2. In total, 821 out of 825 Monte-Carlo cross validation runs were completed. On run 13, 18, 24 and 25 for the ACTG 320 data, the `nn-cox` learner encountered an error during its fitting procedure.

Index of prediction accuracy Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest IPA in 18 out of 33 risk prediction tasks, with an overall mean IPA of 13.2 (Figure 1). Compared to the learner with the second highest mean IPA (`cif-standard`), `aorsf-fast`'s mean was 1.41 points higher, a relative increase of 12.0%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically

[4]For example, when the prediction task was to predict risk of death in the ACTG 320 clinical trial (26 events total), some splits did not leave enough events in the training data to fit complex learners such as the neural network

equivalent expected IPA was 0.99, and the posterior probability of `aorsf-fast` having a superior IPA to other learners ranged from 0.87 (versus `cif-standard`) to >0.999 (versus several other learners; see Figure 2)

Time-dependent concordance statistic Compared to learners that were not oblique RSFs, `aorsf-fast` had the highest time-dependent C-statistic in 8 out of 33 risk prediction tasks, with an overall mean of 77.1 (Figure 3). Compared to the learner with the second highest mean C-statistic (`cif-standard`), `aorsf-fast`'s mean was 0.762 points higher, a relative increase of 0.998%. The posterior probability of `aorsf-fast` and `aorsf-cph` having practically equivalent expected time-dependent C-statistics was 0.99, and the posterior probability of `aorsf-fast` having a superior time-dependent C-statistic versus other learners ranged from 0.24 (versus `cif-standard`) to >0.999 (versus several other learners; see Figure 4)

Computational efficiency Overall, `aorsf-fast` was the second fastest learner, with an expected model development and risk prediction time about 158 milliseconds longer than `glmnet-cox` (Figure 5). Comparing median computing times, `aorsf-fast` was 700.1 times faster than its predecessor, `obliqueRSF-net`. In addition, `aorsf-fast` was 13.6, 1.92, and 4.20 faster than axis based forests grown using the `party`, `ranger`, and `randomForestSRC` packages, respectively.

5.2 Benchmark of variable importance

The aim of this experiment is to evaluate negation VI and similar VI methods based on how well they can discriminate between variables that do or do not have a relationship with a simulated outcome. We consider methods that are intrinsic to the oblique RF (for example, ANOVA VI), those that are intrinsic to the RF (for example, permutation VI), and those that are model-agnostic (for example, SHAP VI). VI methods with unavailable or still developing software were not included.[5]

5.2.1 Variable importance techniques

We compute permutation VI for axis based RSFs using the `randomForestSRC` package. We compute ANOVA VI, negation VI, and permutation VI for oblique RSFs using the `aorsf` package. For ANOVA VI, we applied a p-value threshold of 0.01, following the threshold recommended by Menze et al. [2011]. We compute SHAP VI for boosted tree models using the `xgboost` package, which incorporates the tree SHAP approach proposed by Lundberg et al. [2018].

5.2.2 Variable types

We considered five classes of predictor variables, with each class characterized by its variables' relationship to a right-censored outcome. Specifically,

- *irrelevant* variables had no relationship with the outcome.
- *main effect* variables had a linear relationship to the outcome.
- *non-linear effect* variables had a non-linear relationship to the outcome.
- *combination effect* variables were formed by linear combinations of three other variables. While their combination was linearly related to the outcome, each of the three variables contributing to the combination had no relation to the outcome.
- *interaction effect* variables were related to the outcome by multiplicative interaction with one other variable, which could have been a main effect, non-linear effect, or combination effect variable.

[5]Although the `party` package implements the approach to VI developed by Strobl et al. [2007], the developers of the `party` package note that the implementation of this approach for survival outcomes is "extremely slow and experimental" as of version 1.3.10. Therefore, it is not incorporated in the current simulation study.

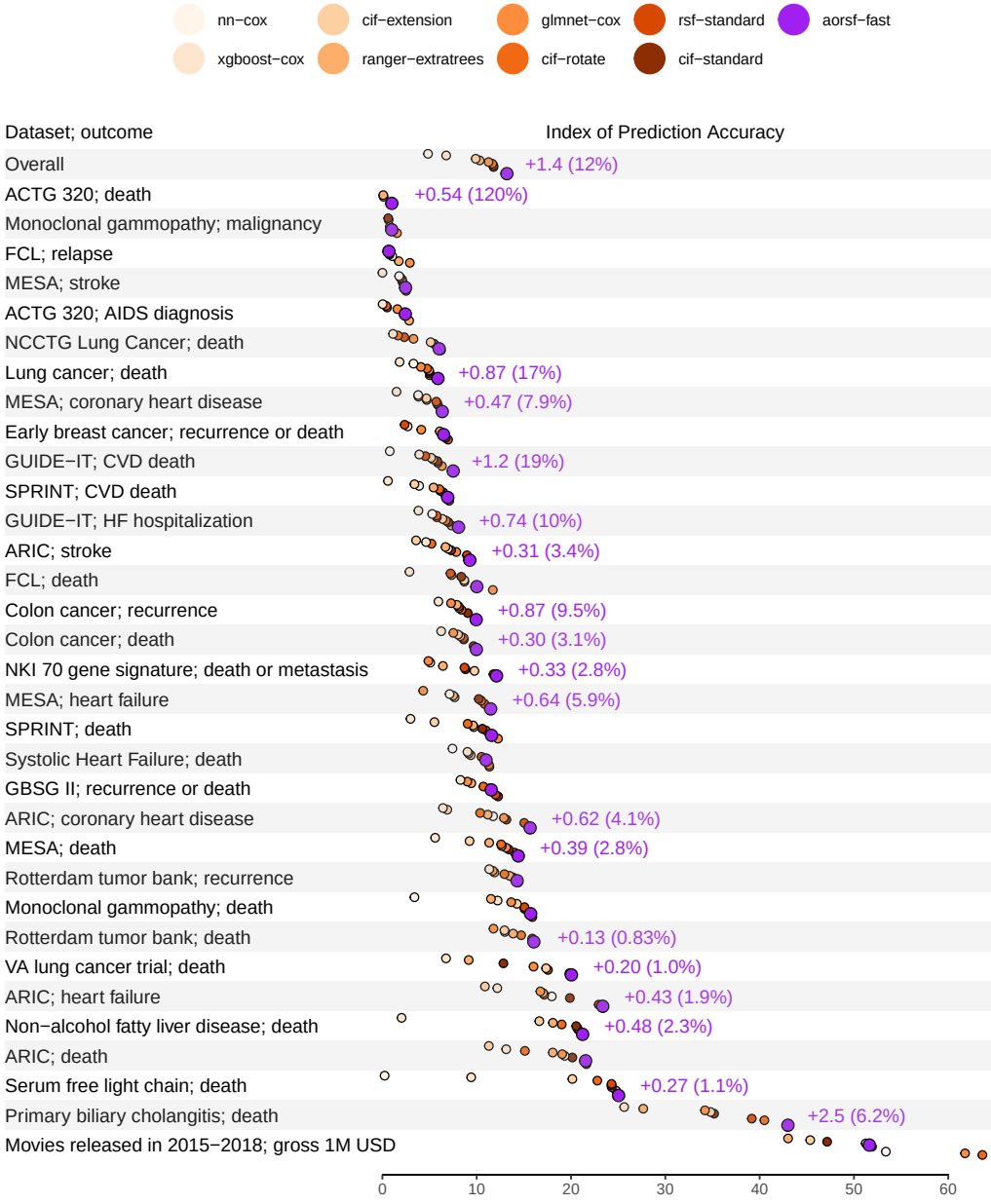

Figure 1: Index of prediction accuracy for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest index of prediction accuracy, showing the absolute and percent improvement over the second best learner. As predicted survival probabilities are not a standard output from `xgboost-aft`, it is not included in this figure.

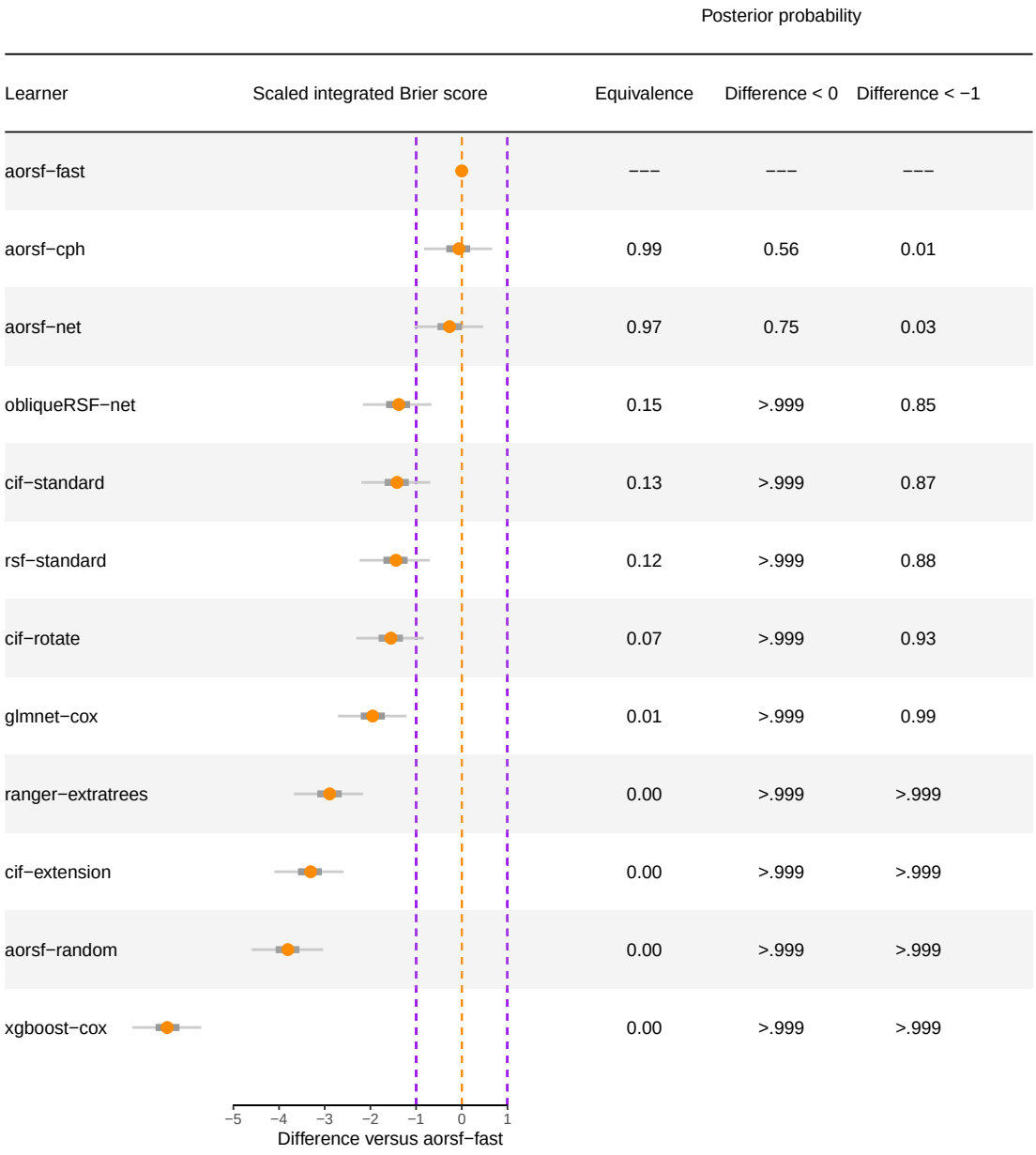

Figure 2: Expected differences in index of prediction accuracy between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

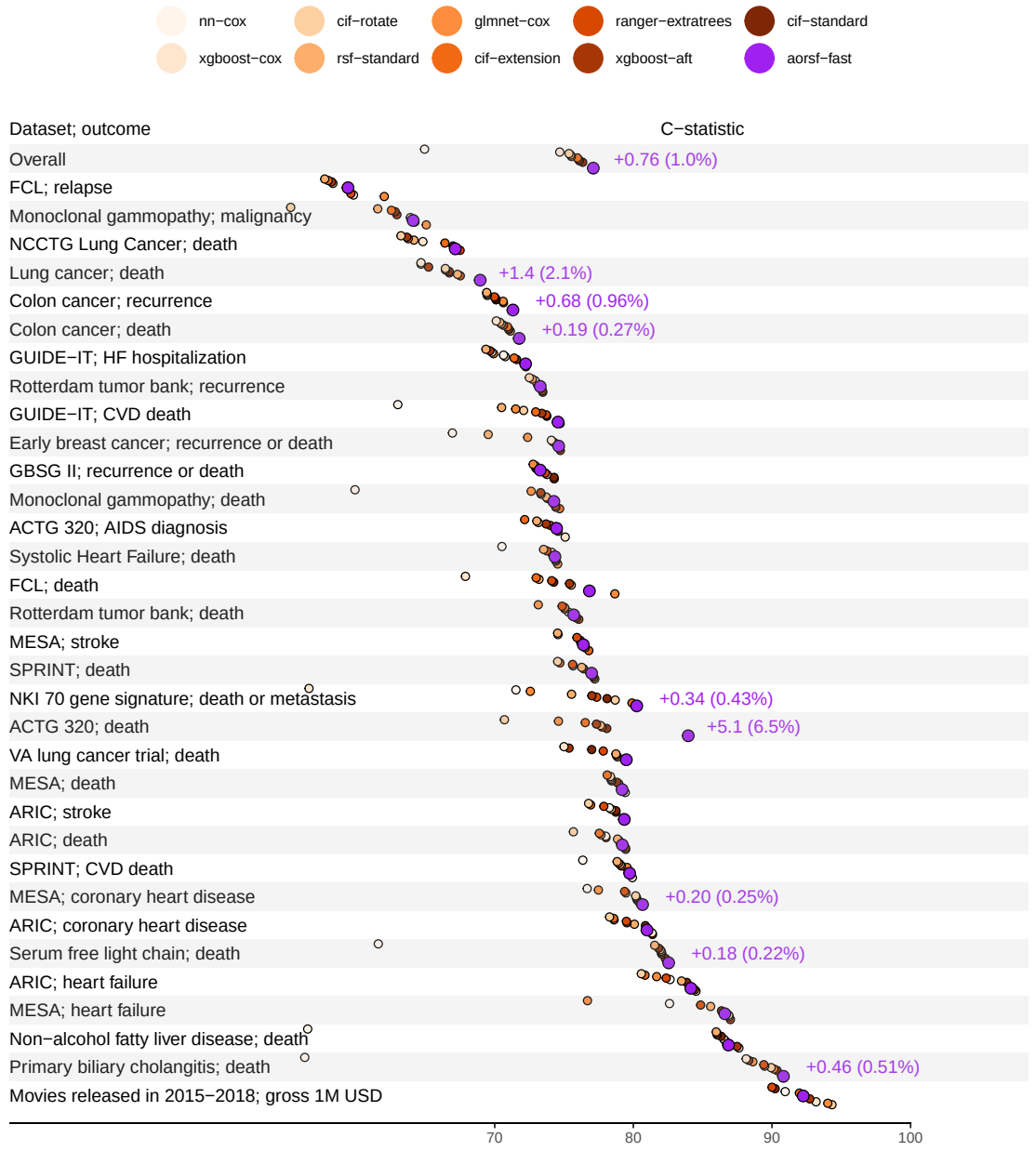

Figure 3: Time-dependent concordance statistic for the accelerated oblique random survival forest and other learning algorithms across multiple risk prediction tasks. Text appears in tasks where the accelerated oblique random survival forest obtained the highest concordance, showing the absolute and percent improvement over the second best learner.

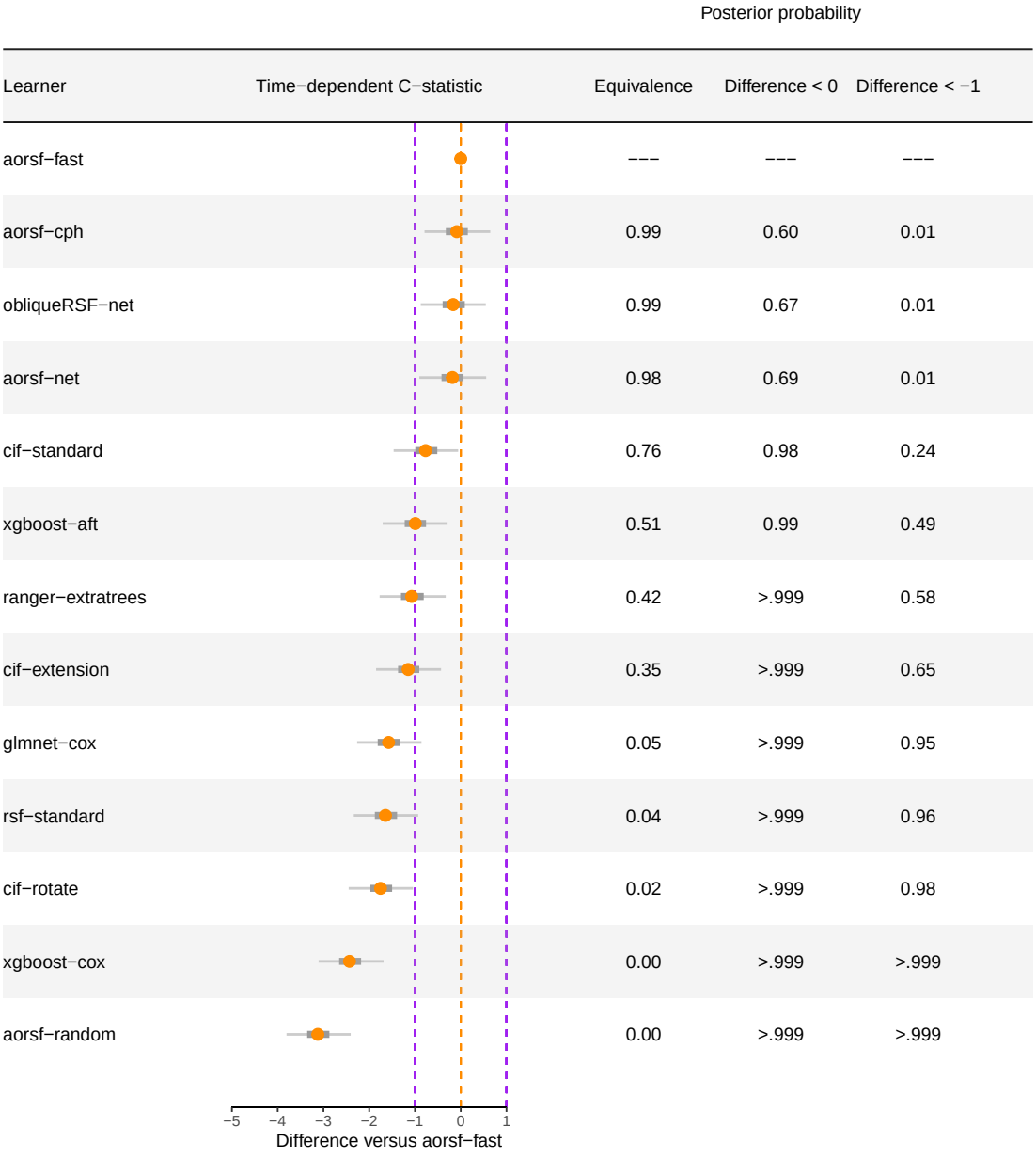

Figure 4: Expected differences in time-dependent concordance statistic between the accelerated oblique random survival forest and other learning algorithms. A region of practical equivalence is shown by purple dotted lines, and a boundary of non-zero difference is shown by an orange dotted line at the origin.

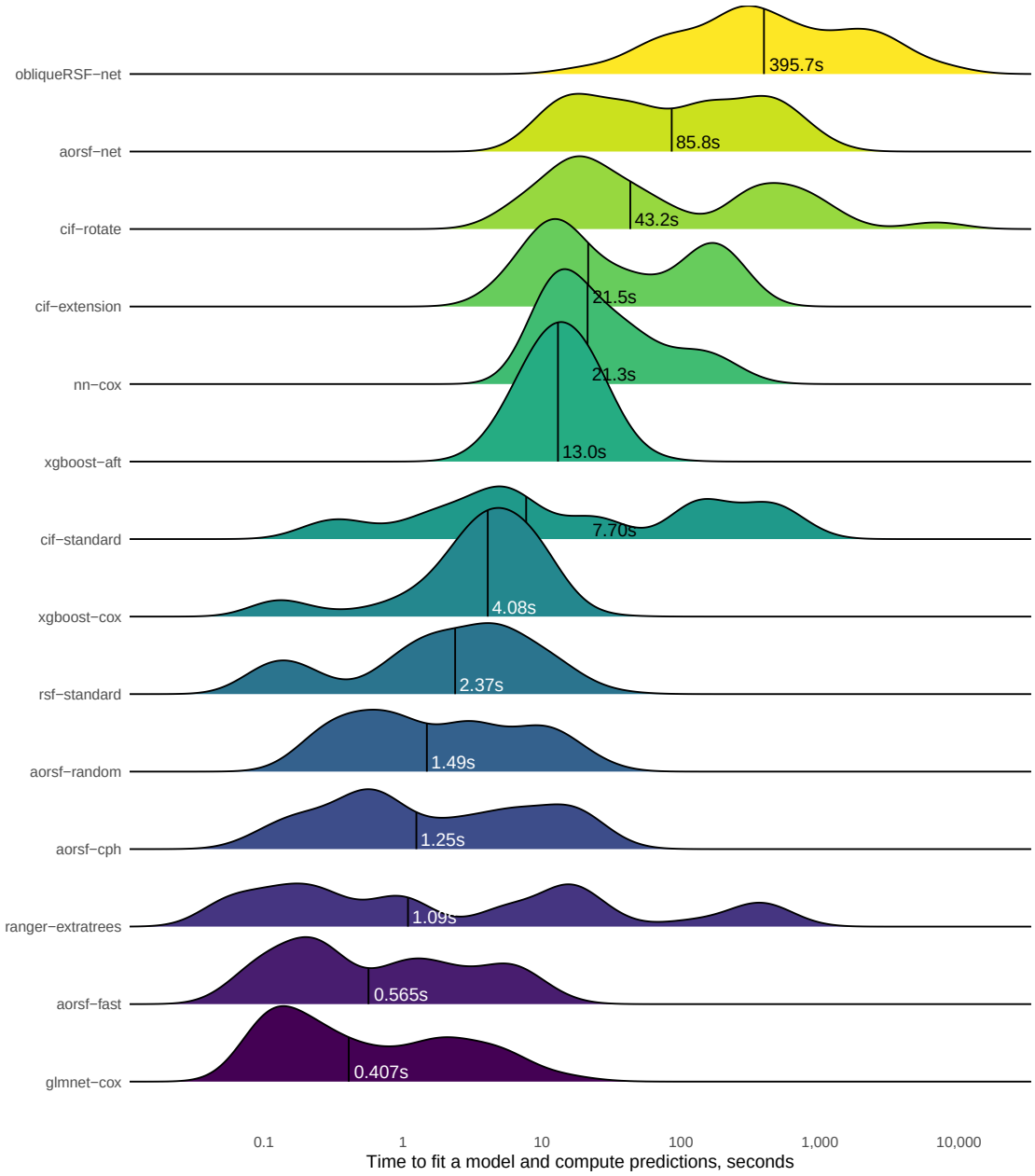

Figure 5: Distribution of time taken to fit a prediction model and compute predicted risk. The median time, in seconds, is printed and annotated for each learner by a vertical line.

5.2.3 Simulated data

We initiated each set of simulated data with a random draw of size n from a p-dimensional multivariate normal distribution, yielding n observations of p predictors. Each of p predictor variables had a mean of zero, standard deviation of 1, and correlation with other predictor variables drawn at random between a lower and upper boundary. A time-to-event outcome with roughly 45% of observations censored was generated using the `simsurv` package. The full predictor matrix (that is, including interactions, non-linear mappings, and combinations) was used to generate the outcome. Interactions, non-linear mappings, and combinations were dropped from the predictor matrix after the outcome was generated so that VI techniques could be evaluated based on their ability to detect these effects.

5.2.4 Parameter specifications

Parameters that varied in the current simulation study included the number of observations (500, 1000, and 2500) and the absolute value of the maximum correlation between predictors (0.3, 0.15, and 0). Parameters that remain fixed throughout the study included the number of predictors in each class (15) and the effect size of each predictor (one standard deviation increase associated with a 64% increase in relative risk). Using this design for simulated data, the proportion of relative risk explained by our covariates ranged from FILL IN to FILL IN, according to Heller's explained relative risk measure [Heller, 2012].

5.2.5 Evaluation of variable importance

We compared VI techniques based on their discrimination (that is, C-statistic) between relevant and irrelevant variables. Specifically, we generated a binary outcome for each predictor variable based on its relevance (that is, the binary outcome is 1 if the variable is relevant, 0 otherwise). Treating VI as if it were a 'prediction' for these binary outcomes yields a C-statistic which may be interpreted as the probability that the VI technique will rank a relevant variable higher than an irrelevant variable [Harrell et al., 1982].

5.2.6 Results

The three techniques that used 'aorsf' to estimate VI were ranked first (`aorsf-negate`; $C = 76.0$), second (`aorsf-anova`; $C = 74.0$), and third (`aorsf-permute`; $C = 73.2$) in overall mean C-statistic across all of the simulation scenarios, with `aorsf-negate` obtaining the highest C-statistic in 26 out of 36 VI tasks (Figure 6). Among the four relevant variable classes, `aorsf-negate` had the highest mean C-statistic for main effects, combination effects, and non-linear effects, with the greatest advantage of using `aorsf-negate` occurring among non-linear and combination variables. Full results from the experiment are provided in Table A.3. Computationally, ANOVA VI was faster than negation and permutation VI, with a median time of 1.27 seconds versus 11.9 and 13.0 seconds, respectively.

6 Discussion

In this paper, we have developed two contributions to the oblique RSF: (1) the accelerated oblique RSF (that is, `aorsf-fast`) and (2) negation VI. Our technique to accelerate the oblique RSF reduces the number of operations required to find linear combinations of inputs using a single iteration of Newton Raphson scoring, while our VI technique directly engages with coefficients in linear combinations of inputs to measure importance of individual variables. In numeric experiments, we found that that `aorsf-fast` is over 500 times faster and just as accurate in risk prediction tasks compared to its predecessor, `obliqueRSF-net`. We also found that negation VI, a technique to estimate VI using the oblique RSF, detected non-linear, combination, and main effects more effectively than three standard methods to estimate VI: permutation, ANOVA, and SHAP VI. Overall, we found that estimating VI using negation instead of ANOVA increased the C-statistic for ranking a relevant variable higher than an irrelevant variable by 1.97, a relative increase of 2.66%.

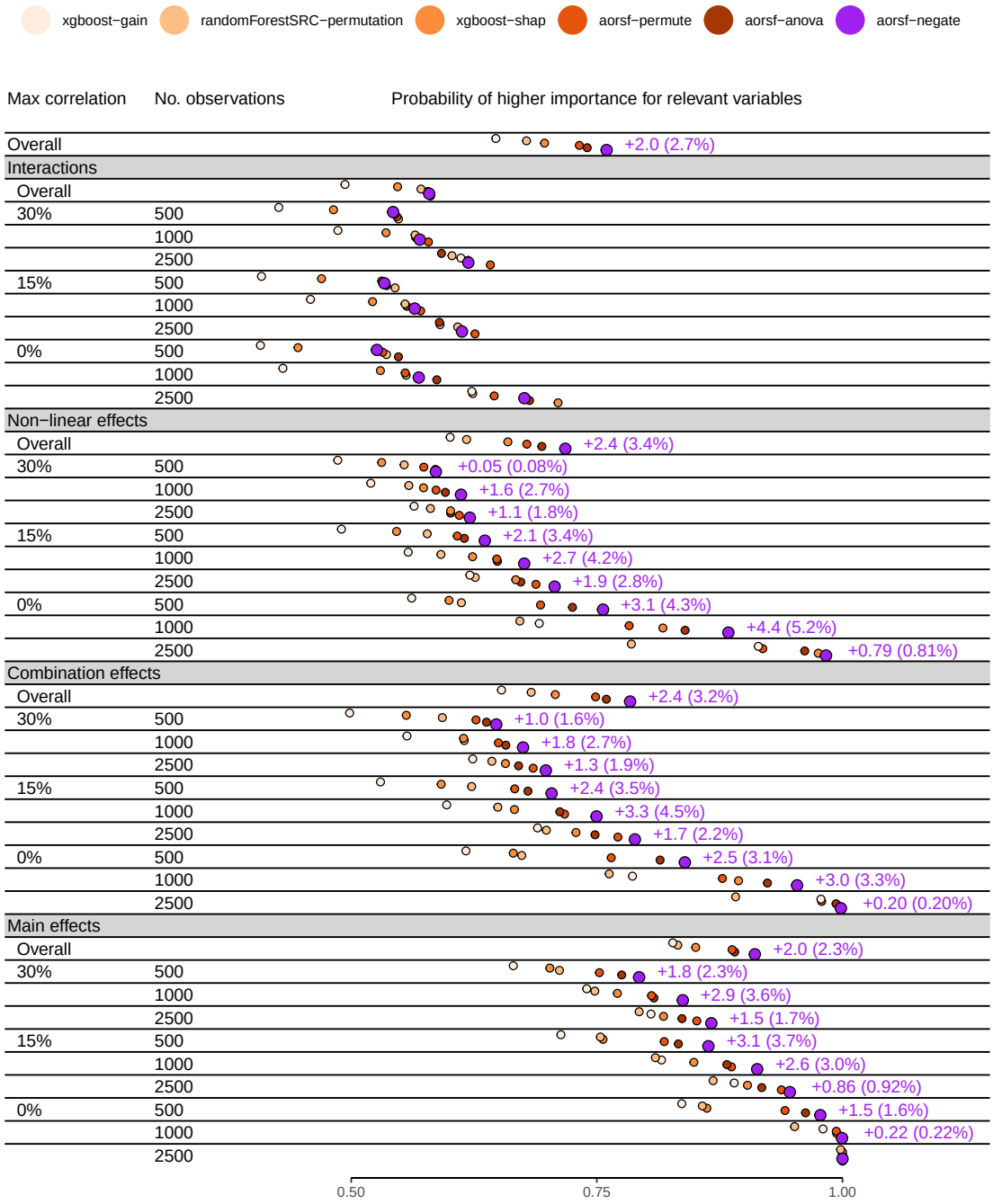

Figure 6: Concordance statistic for assigning higher importance to relevant versus irrelevant variables. Text appears in rows where negation importance obtained the highest concordance, showing absolute and percent improvement over the second best technique.

6.1 Implications of our results

Accurate risk prediction models have the potential to improve healthcare by directing timely interventions to patients who are most likely to benefit. However, prediction models that cannot scale adequately to large databases or cannot be interpreted and explained have no place in clinical practice. The current study advances the oblique RSF, an accurate risk prediction model, towards being accurate, scalable, and interpretable. The improved computational efficiency of the accelerated oblique RSF increases the feasibility of applying oblique RSFs in a wide range of prediction tasks. Faster model evaluation and re-fitting also improve diagnosis and resolution of model-based issues (for example, model calibration deteriorates over time). The introduction of negation VI also advances interpretability. VI is intrinsically linked to model fairness, as it can be used to identify when protected characteristics such as race, religion, and sexuality are inadvertently used (either directly or through correlates of these characteristics) by a prediction model. Since negation VI engages with the coefficients used in linear combinations of variables, a major component of oblique RSFs, it may be more capable of diagnosing unfairness in oblique RSFs compared to permutation importance and model-agnostic VI techniques.

6.2 Limitations and next steps

While the current study advances the oblique RSF towards being scalable and interpretable, there remain several limitations that can be targeted in future studies. The accelerated oblique RSF does not account for competing risks, and biased estimation of incidence may occur when competing risks are ignored. Thus, allowing the oblique RSF to account for competing risks is a high priority for future studies. In addition, missing data are not addressed in the accelerated oblique RSF, and users of the `aorsf` R package are expected to impute missing values prior to model training and testing. However, missing data are common and there are numerous techniques for ensemble tree methods to handle missing data during the tree growing procedure. Thus, a second item of high priority for future studies is to develop and evaluate strategies to handle missing data while growing an oblique RSF. Last, Cui et al. [2017] found that estimating an inverse-probability weighted hazard function at each non-leaf node of a survival tree allows the RSF to converge asymptotically to the true survival function when some variables contribute both to the risk of the event and the risk of censoring, a scenario that is very likely in the analysis of electronic medical records. The accelerated oblique RSF could incorporate this splitting technique by using Newton Raphson scoring to fit a model for the censoring distribution and then a weighted model could be fit to the failure distribution. This final item has the highest priority, as Cui et al. [2017] showed it is a requisite condition for consistency of axis based survival trees in fairly general settings.

Acknowledgements

Research reported in this publication was supported by the Center for Biomedical Informatics, Wake Forest University School of Medicine. The project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR001420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Appendix

Data sources

The "VA lung cancer trial" data [Kalbfleisch and Prentice, 2011] were obtained from the `randomForestSRC` R package [Ishwaran and Kogalur, 2019].

The "Colon cancer" data [Moertel et al., 1995] were obtained from the `survival` R package [Therneau, 2022b].

The "Primary biliary cholangitis" data [Therneau and Grambsch, 2000] were obtained from the `aorsf` R package [Jaeger, 2022].

The "Movies released in 2015-2018" data were obtained from the `censored` R package [Hvitfeldt and Frick].

The "GBSG II" data [Schumacher, 1994] were obtained from the `TH.data` R package [Hothorn, 2022].

The "Systolic Heart Failure" data [Hsich et al., 2011] were obtained from the `randomForestSRC` R package [Ishwaran and Kogalur, 2019].

The "Serum free light chain" data [Dispenzieri et al., 2012, Kyle et al., 2006] were obtained from the `survival` R package [Therneau, 2022b].

The "Non-alcohol fatty liver disease" data [Allen et al., 2018] were obtained from the `survival` R package [Therneau, 2022b].

The "Rotterdam tumor bank" data [Royston and Altman, 2013] were obtained from the `survival` R package [Therneau, 2022b].

The "ACTG 320" data [Hosmer and Lemeshow, 2002] were obtained from the `mlr3proba` R package [Sonabend et al., 2021].

The "GUIDE-IT" data [Felker et al., 2017] were obtained from BioLINCC.

The "Early breast cancer" data [Desmedt et al., 2011, Hatzis et al., 2011, Ternès et al., 2017] were obtained from the `biospear` R package [Ternes et al., 2018].

The "SPRINT" data [SPRINT Research Group, 2015] were obtained from BioLINCC.

The "NKI 70 gene signature" data [Van De Vijver et al., 2002] were obtained from the `OpenML` R package [Casalicchio et al., 2017].

The "Lung cancer" data [Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, 2008] were obtained from the `OpenML` R package [Casalicchio et al., 2017].

The "NCCTG Lung Cancer" data [Loprinzi et al., 1994] were obtained from the `survival` R package [Therneau, 2022b].

The "FCL" data [Pintilie, 2006] were obtained from the `randomForestSRC` R package [Ishwaran and Kogalur, 2019].

The "Monoclonal gammopathy" data [Kyle et al., 2002] were obtained from the `survival` R package [Therneau, 2022b].

The "MESA" data [Bild et al., 2002] were obtained from BioLINCC.

The "ARIC" data [ARIC Investigators, 1989] were obtained from BioLINCC.

A.1: Data sets used for numeric experiments

Label	N observations	N predictors	Outcome	N Events	% Censored
VA lung cancer trial	137	8	Death	128	6.57
Colon cancer	929	12	Recurrence	468	49.6
			Death	452	51.3
Primary biliary cholangitis	276	19	Death	111	59.8
Movies released in 2015-2018	551	46	Gross 1M USD	522	5.26
GBSG II	686	10	Recurrence Or Death	299	56.4
Systolic Heart Failure	2,231	41	Death	726	67.5
Serum free light chain	7,874	10	Death	2,169	72.5
Non-alcohol fatty liver disease	17,549	24	Death	1,364	92.2
Rotterdam tumor bank	2,982	11	Recurrence	1,518	49.1
			Death	1,272	57.3
ACTG 320	1,151	12	AIDS Diagnosis	96	91.7
			Death	26	97.7
GUIDE-IT	894	59	Cardiovascular Death	110	87.7
			Hf Hospitalization	288	67.8
Early breast cancer	614	1,692	Recurrence Or Death	134	78.2
SPRINT	9,361	174	Cardiovascular Death	521	94.4
			Death	1,644	82.4
NKI 70 gene signature	144	77	Death Or Metastasis	48	66.7
Lung cancer	442	24	Death	236	46.6
NCCTG Lung Cancer	228	9	Death	165	27.6
FCL	541	7	Death	76	86.0
			Relapse	272	49.7
			Death	963	30.4

Monoclonal gammopathy	1,384	8	Malignancy	115	91.7
MESA	6,783	48	Heart Failure	339	95.0
			Coronary Heart Disease	439	93.5
			Stroke	292	95.7
			Death	1,297	80.9
ARIC	13,623	41	Heart Failure	2,981	78.1
			Coronary Heart Disease	2,282	83.2
			Stroke	1,323	90.3
			Death	6,662	51.1

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks.

	Performance metric (SD)		Computation time, seconds	
	Scaled Brier	C-Statistic	Model fitting	Risk prediction
Overall				
aorsf-fast	0.132 (0.110)	0.771 (0.073)	0.458	0.107
aorsf-cph	0.131 (0.110)	0.770 (0.072)	1.145	0.087
aorsf-net	0.129 (0.113)	0.769 (0.073)	85.734	0.087
obliqueRSF-net	0.118 (0.085)	0.769 (0.073)	356.969	17.203
cif-standard	0.118 (0.098)	0.763 (0.072)	2.035	5.650
rsf-standard	0.118 (0.115)	0.755 (0.077)	2.157	0.194
cif-rotate	0.117 (0.126)	0.754 (0.083)	35.290	6.347
glmnet-cox	0.113 (0.121)	0.755 (0.079)	0.405	0.003
ranger-extratrees	0.103 (0.086)	0.760 (0.069)	0.716	0.848
cif-extension	0.099 (0.094)	0.760 (0.074)	15.496	6.211
aorsf-random	0.094 (0.081)	0.740 (0.065)	1.339	0.086
xgboost-cox	0.068 (0.102)	0.747 (0.096)	4.073	0.004
nn-cox	0.048 (0.111)	0.649 (0.139)	16.050	1.501
xgboost-aft	—	0.761 (0.078)	13.035	0.007
ACTG 320; AIDS diagnosis, n = 1151, p = 12				
aorsf-random	0.028 (0.022)	0.748 (0.038)	0.539	0.040
ranger-extratrees	0.028 (0.017)	0.740 (0.036)	0.064	0.146
obliqueRSF-net	0.027 (0.022)	0.746 (0.038)	25.980	15.236
aorsf-cph	0.025 (0.029)	0.751 (0.042)	0.471	0.033
cif-standard	0.024 (0.031)	0.744 (0.040)	1.771	4.657
aorsf-fast	0.024 (0.028)	0.745 (0.044)	0.198	0.033
cif-extension	0.023 (0.015)	0.722 (0.038)	9.848	4.083
aorsf-net	0.019 (0.034)	0.745 (0.042)	18.451	0.035
glmnet-cox	0.016 (0.030)	0.746 (0.037)	0.179	0.003
rsf-standard	0.005 (0.041)	0.730 (0.042)	0.192	0.061
cif-rotate	0.004 (0.040)	0.731 (0.038)	16.095	3.888
xgboost-cox	0.000 (0.044)	0.751 (0.033)	3.704	0.003
nn-cox	-0.001 (0.008)	0.549 (0.125)	11.075	0.668
xgboost-aft	—	0.737 (0.035)	10.015	0.006
ACTG 320; death, n = 1151, p = 12				
aorsf-fast	0.010 (0.022)	0.840 (0.054)	0.086	0.020
obliqueRSF-net	0.007 (0.011)	0.823 (0.052)	8.869	11.004
aorsf-cph	0.007 (0.018)	0.821 (0.060)	0.414	0.020
aorsf-random	0.005 (0.014)	0.788 (0.074)	0.278	0.023
cif-extension	0.001 (0.020)	0.765 (0.066)	8.436	3.493
ranger-extratrees	0.001 (0.019)	0.777 (0.069)	0.044	0.138
xgboost-cox	-0.004 (0.004)	0.500 (0.000)	0.119	0.002
nn-cox	-0.004 (0.004)	0.508 (0.119)	10.867	0.518
cif-standard	-0.005 (0.025)	0.781 (0.062)	1.726	4.648
aorsf-net	-0.005 (0.032)	0.806 (0.067)	14.625	0.023
rsf-standard	-0.031 (0.051)	0.776 (0.073)	0.090	0.035
cif-rotate	-0.037 (0.049)	0.707 (0.090)	13.920	3.444

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
glmnet-cox	-0.065 (0.095)	0.746 (0.098)	0.281	0.002
xgboost-aft	—	0.773 (0.071)	9.190	0.005
ARIC; coronary heart disease, n = 13623, p = 41				
aorsf-fast	0.157 (0.007)	0.810 (0.007)	4.640	1.320
aorsf-cph	0.153 (0.006)	0.809 (0.007)	14.503	1.347
aorsf-net	0.152 (0.006)	0.809 (0.007)	533.804	1.437
rsf-standard	0.150 (0.007)	0.801 (0.007)	9.407	0.991
obliqueRSF-net	0.143 (0.005)	0.811 (0.008)	2830.398	356.908
cif-standard	0.132 (0.005)	0.809 (0.007)	71.835	371.102
glmnet-cox	0.129 (0.011)	0.795 (0.008)	1.473	0.011
nn-cox	0.117 (0.029)	0.786 (0.046)	51.356	84.715
ranger-extratrees	0.112 (0.005)	0.795 (0.009)	293.864	61.837
cif-rotate	0.104 (0.004)	0.783 (0.009)	569.289	67.436
aorsf-random	0.098 (0.005)	0.772 (0.008)	11.694	1.330
cif-extension	0.069 (0.002)	0.786 (0.009)	169.206	56.091
xgboost-cox	0.064 (0.017)	0.814 (0.006)	8.544	0.015
xgboost-aft	—	0.814 (0.006)	23.781	0.013
ARIC; death, n = 13623, p = 41				
aorsf-net	0.217 (0.006)	0.792 (0.004)	932.212	2.377
rsf-standard	0.216 (0.006)	0.789 (0.004)	14.727	1.274
aorsf-cph	0.216 (0.006)	0.792 (0.004)	22.922	2.179
aorsf-fast	0.215 (0.007)	0.792 (0.004)	7.641	2.214
obliqueRSF-net	0.207 (0.005)	0.791 (0.004)	7329.229	331.164
cif-standard	0.201 (0.004)	0.790 (0.004)	72.253	384.744
nn-cox	0.193 (0.011)	0.780 (0.005)	103.296	92.992
glmnet-cox	0.191 (0.015)	0.777 (0.007)	2.256	0.011
ranger-extratrees	0.181 (0.004)	0.780 (0.005)	332.296	70.372
cif-rotate	0.151 (0.007)	0.757 (0.006)	589.958	66.749
xgboost-cox	0.131 (0.012)	0.794 (0.004)	11.220	0.015
aorsf-random	0.130 (0.005)	0.734 (0.005)	20.175	2.057
cif-extension	0.113 (0.002)	0.775 (0.005)	180.483	54.092
xgboost-aft	—	0.794 (0.004)	28.692	0.014
ARIC; heart failure, n = 13623, p = 41				
aorsf-fast	0.234 (0.006)	0.841 (0.005)	5.268	2.193
rsf-standard	0.229 (0.006)	0.835 (0.005)	11.238	1.079
aorsf-cph	0.229 (0.006)	0.841 (0.005)	16.864	2.308
aorsf-net	0.228 (0.006)	0.841 (0.005)	623.727	1.685
obliqueRSF-net	0.212 (0.005)	0.841 (0.005)	3737.617	320.616
cif-standard	0.199 (0.005)	0.839 (0.005)	71.854	391.419
nn-cox	0.180 (0.017)	0.826 (0.007)	56.056	96.125
cif-rotate	0.172 (0.006)	0.806 (0.007)	589.494	69.448
ranger-extratrees	0.170 (0.004)	0.824 (0.005)	387.176	77.513
glmnet-cox	0.167 (0.044)	0.817 (0.018)	2.141	0.011
aorsf-random	0.139 (0.005)	0.789 (0.006)	14.001	1.519
xgboost-cox	0.122 (0.017)	0.845 (0.005)	10.926	0.015

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
cif-extension	0.109 (0.003)	0.808 (0.006)	176.907	57.216
xgboost-aft	—	0.844 (0.005)	24.549	0.013
ARIC; stroke, n = 13623, p = 41				
aorsf-fast	0.093 (0.004)	0.793 (0.007)	4.179	1.122
aorsf-net	0.090 (0.004)	0.792 (0.007)	371.989	1.175
aorsf-cph	0.090 (0.004)	0.792 (0.007)	13.462	1.149
rsf-standard	0.090 (0.006)	0.784 (0.006)	8.991	0.984
obliqueRSF-net	0.082 (0.003)	0.791 (0.007)	1869.359	387.014
glmnet-cox	0.078 (0.004)	0.787 (0.007)	1.673	0.011
cif-standard	0.073 (0.003)	0.787 (0.007)	73.989	364.863
nn-cox	0.070 (0.012)	0.783 (0.010)	32.099	82.720
ranger-extratrees	0.067 (0.003)	0.779 (0.008)	251.130	50.542
aorsf-random	0.059 (0.004)	0.750 (0.008)	9.712	1.132
cif-rotate	0.052 (0.003)	0.768 (0.009)	586.962	68.410
xgboost-cox	0.046 (0.014)	0.794 (0.006)	7.316	0.015
cif-extension	0.036 (0.002)	0.769 (0.009)	169.866	53.493
xgboost-aft	—	0.793 (0.006)	20.363	0.013
Colon cancer; death, n = 929, p = 12				
aorsf-fast	0.100 (0.014)	0.718 (0.012)	0.239	0.053
aorsf-cph	0.099 (0.014)	0.717 (0.011)	0.638	0.052
cif-standard	0.097 (0.013)	0.710 (0.012)	0.745	3.685
aorsf-net	0.096 (0.014)	0.717 (0.012)	49.140	0.047
aorsf-random	0.096 (0.010)	0.716 (0.011)	0.919	0.044
obliqueRSF-net	0.089 (0.006)	0.717 (0.012)	252.190	16.726
cif-rotate	0.086 (0.017)	0.705 (0.014)	13.242	3.448
rsf-standard	0.086 (0.019)	0.704 (0.011)	1.901	0.150
ranger-extratrees	0.083 (0.007)	0.710 (0.012)	0.560	0.248
cif-extension	0.080 (0.006)	0.709 (0.011)	8.397	3.363
glmnet-cox	0.075 (0.016)	0.711 (0.019)	0.128	0.003
xgboost-cox	0.062 (0.013)	0.701 (0.013)	3.217	0.003
nn-cox	-0.003 (0.003)	0.508 (0.033)	12.769	1.106
xgboost-aft	—	0.706 (0.013)	10.290	0.006
Colon cancer; recurrence, n = 929, p = 12				
aorsf-fast	0.099 (0.017)	0.713 (0.016)	0.225	0.050
aorsf-cph	0.098 (0.017)	0.712 (0.015)	0.629	0.050
aorsf-net	0.095 (0.018)	0.713 (0.017)	49.866	0.048
aorsf-random	0.091 (0.013)	0.706 (0.013)	0.926	0.043
cif-standard	0.091 (0.016)	0.701 (0.017)	0.717	3.703
obliqueRSF-net	0.087 (0.009)	0.711 (0.015)	249.790	16.701
cif-rotate	0.084 (0.020)	0.694 (0.017)	13.154	3.977
cif-extension	0.081 (0.009)	0.706 (0.017)	8.446	4.066
rsf-standard	0.081 (0.020)	0.694 (0.015)	1.831	0.153
ranger-extratrees	0.079 (0.011)	0.700 (0.016)	0.636	0.254
glmnet-cox	0.073 (0.018)	0.706 (0.024)	0.125	0.003
xgboost-cox	0.059 (0.011)	0.695 (0.018)	3.136	0.003

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	-0.005 (0.006)	0.531 (0.054)	13.107	0.969
xgboost-aft	—	0.701 (0.019)	11.227	0.006
Early breast cancer; recurrence or death, n = 614, p = 1692				
obliqueRSF-net	0.072 (0.023)	0.751 (0.029)	1901.883	14.017
cif-rotate	0.070 (0.018)	0.747 (0.027)	6452.563	360.067
cif-standard	0.067 (0.019)	0.747 (0.030)	9.660	4.247
aorsf-cph	0.067 (0.029)	0.747 (0.026)	1.287	0.185
aorsf-fast	0.065 (0.028)	0.746 (0.026)	0.724	0.177
cif-extension	0.064 (0.016)	0.746 (0.028)	44.047	6.554
ranger-extratrees	0.061 (0.022)	0.742 (0.031)	0.228	0.169
glmnet-cox	0.041 (0.032)	0.724 (0.036)	5.780	0.006
aorsf-random	0.027 (0.015)	0.696 (0.038)	1.260	0.175
xgboost-cox	0.027 (0.034)	0.741 (0.030)	2.258	0.007
rsf-standard	0.024 (0.037)	0.695 (0.033)	0.366	0.743
aorsf-net	0.012 (0.062)	0.740 (0.025)	465.845	0.175
nn-cox	-0.006 (0.043)	0.669 (0.075)	19.374	1.808
xgboost-aft	—	0.744 (0.027)	9.708	0.010
FCL; death, n = 541, p = 7				
glmnet-cox	0.117 (0.028)	0.787 (0.037)	0.098	0.002
aorsf-cph	0.100 (0.039)	0.769 (0.033)	0.169	0.018
aorsf-fast	0.100 (0.038)	0.768 (0.033)	0.090	0.018
aorsf-net	0.097 (0.040)	0.760 (0.034)	12.619	0.018
obliqueRSF-net	0.089 (0.027)	0.758 (0.036)	97.158	5.447
cif-rotate	0.087 (0.048)	0.755 (0.027)	6.083	1.964
cif-extension	0.087 (0.036)	0.730 (0.034)	5.256	2.514
aorsf-random	0.087 (0.029)	0.757 (0.032)	0.256	0.018
cif-standard	0.084 (0.038)	0.743 (0.036)	0.347	1.132
ranger-extratrees	0.073 (0.016)	0.741 (0.037)	0.043	0.078
rsf-standard	0.072 (0.048)	0.732 (0.034)	0.156	0.040
xgboost-cox	0.029 (0.050)	0.679 (0.121)	0.266	0.002
nn-cox	-0.004 (0.014)	0.549 (0.110)	11.152	0.406
xgboost-aft	—	0.754 (0.038)	6.738	0.005
FCL; relapse, n = 541, p = 7				
glmnet-cox	0.029 (0.017)	0.620 (0.024)	0.100	0.002
ranger-extratrees	0.017 (0.016)	0.596 (0.025)	0.041	0.081
obliqueRSF-net	0.013 (0.016)	0.593 (0.024)	217.838	6.232
aorsf-random	0.013 (0.018)	0.595 (0.024)	0.394	0.021
xgboost-cox	0.010 (0.016)	0.598 (0.031)	1.331	0.002
cif-standard	0.008 (0.021)	0.594 (0.023)	0.315	1.056
aorsf-cph	0.007 (0.020)	0.595 (0.026)	0.262	0.021
aorsf-fast	0.007 (0.019)	0.594 (0.025)	0.122	0.021
aorsf-net	0.006 (0.020)	0.592 (0.026)	18.761	0.022
cif-extension	-0.005 (0.023)	0.580 (0.028)	6.164	2.309
nn-cox	-0.008 (0.023)	0.526 (0.055)	12.031	0.466
cif-rotate	-0.012 (0.025)	0.583 (0.030)	7.411	2.486

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
rsf-standard	-0.026 (0.032)	0.577 (0.024)	1.081	0.088
xgboost-aft	—	0.582 (0.034)	6.223	0.005
GBSG II; recurrence or death, n = 686, p = 10				
obliqueRSF-net	0.124 (0.016)	0.746 (0.017)	286.149	6.874
cif-standard	0.123 (0.020)	0.743 (0.020)	0.507	2.289
rsf-standard	0.120 (0.023)	0.738 (0.019)	1.437	0.115
aorsf-net	0.120 (0.024)	0.738 (0.020)	36.113	0.038
aorsf-cph	0.120 (0.025)	0.736 (0.018)	0.411	0.038
aorsf-fast	0.115 (0.024)	0.733 (0.017)	0.168	0.037
cif-extension	0.114 (0.017)	0.743 (0.019)	7.560	2.919
cif-rotate	0.107 (0.023)	0.729 (0.017)	11.241	3.023
aorsf-random	0.104 (0.023)	0.724 (0.025)	0.754	0.036
ranger-extratrees	0.094 (0.018)	0.736 (0.025)	0.102	0.143
glmnet-cox	0.090 (0.019)	0.728 (0.021)	0.106	0.002
xgboost-cox	0.083 (0.017)	0.730 (0.020)	2.538	0.003
nn-cox	-0.007 (0.012)	0.509 (0.061)	11.658	0.746
xgboost-aft	—	0.729 (0.021)	10.625	0.006
GUIDE-IT; CVD death, n = 894, p = 59				
aorsf-fast	0.075 (0.018)	0.746 (0.027)	0.152	0.036
aorsf-net	0.074 (0.019)	0.743 (0.027)	29.058	0.041
aorsf-cph	0.071 (0.018)	0.741 (0.028)	0.403	0.037
glmnet-cox	0.063 (0.041)	0.715 (0.091)	0.502	0.002
obliqueRSF-net	0.063 (0.013)	0.741 (0.023)	221.806	11.417
cif-rotate	0.059 (0.016)	0.721 (0.025)	35.190	4.947
cif-standard	0.058 (0.014)	0.738 (0.022)	1.564	3.334
ranger-extratrees	0.054 (0.013)	0.737 (0.029)	0.118	0.176
cif-extension	0.052 (0.011)	0.730 (0.022)	13.507	5.567
rsf-standard	0.046 (0.023)	0.705 (0.025)	0.195	0.064
xgboost-cox	0.039 (0.051)	0.747 (0.020)	3.831	0.003
aorsf-random	0.033 (0.012)	0.695 (0.030)	0.536	0.039
nn-cox	0.008 (0.018)	0.630 (0.123)	12.211	0.596
xgboost-aft	—	0.734 (0.020)	11.525	0.006
GUIDE-IT; HF hospitalization, n = 894, p = 59				
aorsf-net	0.082 (0.017)	0.722 (0.023)	53.095	0.058
aorsf-cph	0.081 (0.018)	0.722 (0.023)	0.659	0.052
aorsf-fast	0.081 (0.019)	0.722 (0.025)	0.232	0.052
ranger-extratrees	0.073 (0.010)	0.722 (0.022)	0.137	0.198
obliqueRSF-net	0.073 (0.010)	0.721 (0.023)	389.465	9.365
cif-standard	0.070 (0.010)	0.716 (0.023)	1.468	3.315
cif-rotate	0.067 (0.019)	0.708 (0.029)	41.655	5.185
cif-extension	0.064 (0.009)	0.714 (0.022)	15.396	6.113
glmnet-cox	0.058 (0.020)	0.699 (0.025)	0.416	0.003
rsf-standard	0.058 (0.022)	0.694 (0.026)	1.515	0.121
nn-cox	0.053 (0.028)	0.706 (0.032)	12.912	0.599
aorsf-random	0.049 (0.010)	0.682 (0.023)	0.914	0.054

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
xgboost-cox	0.038 (0.017)	0.698 (0.027)	2.955	0.003
xgboost-aft	—	0.697 (0.025)	11.409	0.006
Lung cancer; death, n = 442, p = 24				
aorsf-cph	0.063 (0.032)	0.691 (0.019)	0.316	0.030
aorsf-net	0.061 (0.030)	0.685 (0.019)	31.742	0.031
aorsf-fast	0.059 (0.033)	0.690 (0.019)	0.125	0.030
obliqueRSF-net	0.058 (0.020)	0.678 (0.020)	294.626	2.961
cif-extension	0.050 (0.018)	0.667 (0.019)	9.104	3.317
rsf-standard	0.050 (0.035)	0.673 (0.023)	0.941	0.073
cif-standard	0.050 (0.023)	0.667 (0.022)	0.315	0.846
ranger-extratrees	0.049 (0.016)	0.675 (0.019)	0.037	0.065
cif-rotate	0.047 (0.026)	0.664 (0.021)	17.129	3.089
glmnet-cox	0.041 (0.024)	0.664 (0.034)	0.123	0.002
aorsf-random	0.040 (0.022)	0.651 (0.023)	0.538	0.026
nn-cox	0.033 (0.025)	0.647 (0.029)	12.472	0.368
xgboost-cox	0.018 (0.019)	0.647 (0.027)	1.541	0.002
xgboost-aft	—	0.652 (0.026)	8.410	0.006
MESA; coronary heart disease, n = 6785, p = 48				
aorsf-fast	0.063 (0.010)	0.807 (0.011)	1.219	0.352
aorsf-net	0.062 (0.010)	0.805 (0.012)	176.990	0.387
obliqueRSF-net	0.062 (0.008)	0.808 (0.012)	488.662	265.171
aorsf-cph	0.060 (0.010)	0.801 (0.012)	5.393	0.382
cif-standard	0.059 (0.007)	0.803 (0.013)	24.446	98.767
cif-rotate	0.058 (0.009)	0.802 (0.013)	284.668	37.340
rsf-standard	0.057 (0.012)	0.795 (0.013)	3.426	1.202
ranger-extratrees	0.047 (0.004)	0.794 (0.011)	7.360	6.260
cif-extension	0.047 (0.003)	0.805 (0.013)	98.278	28.773
glmnet-cox	0.038 (0.017)	0.775 (0.016)	4.907	0.007
nn-cox	0.038 (0.017)	0.767 (0.021)	18.380	16.712
aorsf-random	0.031 (0.005)	0.735 (0.015)	2.927	0.397
xgboost-cox	0.015 (0.028)	0.802 (0.013)	4.763	0.009
xgboost-aft	—	0.802 (0.012)	18.127	0.009
MESA; death, n = 6793, p = 48				
aorsf-net	0.144 (0.008)	0.792 (0.009)	318.785	0.546
aorsf-fast	0.144 (0.009)	0.792 (0.009)	1.694	0.506
aorsf-cph	0.143 (0.008)	0.791 (0.009)	6.893	0.523
rsf-standard	0.140 (0.008)	0.784 (0.009)	4.818	0.502
obliqueRSF-net	0.139 (0.007)	0.791 (0.009)	1176.365	156.871
cif-standard	0.134 (0.007)	0.788 (0.009)	23.671	101.207
glmnet-cox	0.131 (0.026)	0.789 (0.012)	1.560	0.007
nn-cox	0.127 (0.020)	0.784 (0.016)	29.272	17.344
cif-rotate	0.126 (0.007)	0.783 (0.010)	319.953	37.290
ranger-extratrees	0.113 (0.004)	0.784 (0.008)	9.060	6.474
cif-extension	0.092 (0.003)	0.781 (0.009)	111.068	32.105
aorsf-random	0.068 (0.005)	0.725 (0.008)	5.608	0.565

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
xgboost-cox	0.056 (0.029)	0.794 (0.009)	8.021	0.009
xgboost-aft	—	0.793 (0.009)	20.287	0.010
MESA; heart failure, n = 6785, p = 48				
aorsf-fast	0.115 (0.010)	0.866 (0.013)	1.094	0.311
aorsf-net	0.114 (0.011)	0.863 (0.013)	149.605	0.339
aorsf-cph	0.109 (0.011)	0.858 (0.014)	4.842	0.326
rsf-standard	0.108 (0.012)	0.856 (0.012)	3.125	1.151
obliqueRSF-net	0.108 (0.008)	0.869 (0.011)	393.285	338.322
cif-rotate	0.105 (0.010)	0.869 (0.013)	260.914	38.177
cif-standard	0.102 (0.009)	0.864 (0.013)	24.344	101.534
cif-extension	0.077 (0.005)	0.864 (0.011)	94.628	30.361
ranger-extratrees	0.075 (0.005)	0.849 (0.015)	7.492	6.923
nn-cox	0.071 (0.024)	0.826 (0.016)	15.797	17.312
aorsf-random	0.064 (0.006)	0.795 (0.014)	2.520	0.369
glmnet-cox	0.043 (0.044)	0.767 (0.139)	3.777	0.006
xgboost-cox	-0.008 (0.019)	0.869 (0.011)	6.764	0.009
xgboost-aft	—	0.870 (0.012)	18.620	0.010
MESA; stroke, n = 6783, p = 48				
obliqueRSF-net	0.025 (0.004)	0.767 (0.016)	357.039	299.786
cif-rotate	0.025 (0.004)	0.764 (0.017)	268.402	37.899
cif-standard	0.025 (0.004)	0.762 (0.017)	23.450	98.166
aorsf-fast	0.025 (0.006)	0.764 (0.016)	1.072	0.307
aorsf-net	0.024 (0.006)	0.759 (0.017)	139.617	0.333
aorsf-cph	0.023 (0.005)	0.758 (0.016)	4.266	0.316
ranger-extratrees	0.022 (0.003)	0.759 (0.016)	7.610	6.738
glmnet-cox	0.021 (0.009)	0.765 (0.017)	3.876	0.007
cif-extension	0.021 (0.002)	0.768 (0.017)	94.913	29.730
rsf-standard	0.019 (0.009)	0.745 (0.018)	3.221	1.242
nn-cox	0.018 (0.007)	0.746 (0.028)	16.152	17.979
aorsf-random	0.013 (0.003)	0.714 (0.022)	2.420	0.343
xgboost-cox	0.000 (0.025)	0.762 (0.018)	4.347	0.008
xgboost-aft	—	0.764 (0.015)	16.257	0.009
Monoclonal gammopathy; death, n = 1384, p = 8				
cif-rotate	0.159 (0.019)	0.744 (0.014)	15.123	4.858
aorsf-cph	0.158 (0.016)	0.743 (0.011)	1.154	0.084
aorsf-fast	0.157 (0.016)	0.743 (0.011)	0.408	0.088
aorsf-net	0.155 (0.016)	0.741 (0.011)	85.428	0.085
obliqueRSF-net	0.155 (0.013)	0.743 (0.011)	232.467	12.924
cif-standard	0.151 (0.015)	0.738 (0.012)	1.537	5.829
rsf-standard	0.151 (0.017)	0.737 (0.011)	2.281	0.194
aorsf-random	0.146 (0.013)	0.735 (0.011)	1.757	0.084
cif-extension	0.143 (0.009)	0.747 (0.013)	11.222	4.744
glmnet-cox	0.137 (0.021)	0.726 (0.014)	0.136	0.003
xgboost-cox	0.122 (0.012)	0.733 (0.012)	3.919	0.003
ranger-extratrees	0.115 (0.005)	0.744 (0.012)	0.064	0.181

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	0.034 (0.057)	0.599 (0.112)	16.529	0.689
xgboost-aft	—	0.733 (0.013)	12.094	0.006
Monoclonal gammopathy; malignancy, n = 1384, p = 8				
glmnet-cox	0.015 (0.011)	0.651 (0.055)	0.118	0.003
aorsf-cph	0.011 (0.013)	0.644 (0.036)	0.574	0.040
aorsf-fast	0.010 (0.014)	0.641 (0.036)	0.201	0.041
ranger-extratrees	0.008 (0.006)	0.642 (0.030)	0.095	0.177
cif-extension	0.008 (0.010)	0.625 (0.028)	9.141	4.788
obliqueRSF-net	0.007 (0.010)	0.628 (0.033)	41.856	16.910
aorsf-net	0.007 (0.014)	0.641 (0.034)	22.366	0.041
aorsf-random	0.007 (0.013)	0.633 (0.033)	0.517	0.040
xgboost-cox	0.007 (0.017)	0.639 (0.039)	1.783	0.003
cif-standard	0.006 (0.011)	0.628 (0.033)	1.689	5.603
nn-cox	-0.003 (0.005)	0.510 (0.032)	11.229	0.597
rsf-standard	-0.009 (0.018)	0.616 (0.036)	0.824	0.073
cif-rotate	-0.024 (0.023)	0.553 (0.035)	12.975	4.603
xgboost-aft	—	0.629 (0.039)	10.748	0.006
Movies released in 2015-2018; gross 1M USD, n = 551, p = 46				
cif-rotate	0.636 (0.024)	0.943 (0.007)	19.879	3.571
glmnet-cox	0.618 (0.034)	0.940 (0.009)	0.204	0.002
nn-cox	0.534 (0.072)	0.910 (0.027)	17.266	0.663
aorsf-net	0.530 (0.028)	0.928 (0.010)	50.720	0.043
aorsf-cph	0.522 (0.024)	0.925 (0.011)	0.788	0.041
rsf-standard	0.519 (0.022)	0.922 (0.010)	1.631	0.106
aorsf-fast	0.516 (0.027)	0.923 (0.012)	0.214	0.041
xgboost-cox	0.512 (0.029)	0.932 (0.009)	13.972	0.004
cif-standard	0.472 (0.029)	0.902 (0.018)	0.453	1.910
cif-extension	0.454 (0.025)	0.920 (0.013)	8.854	3.924
ranger-extratrees	0.430 (0.025)	0.900 (0.019)	0.049	0.103
obliqueRSF-net	0.319 (0.022)	0.909 (0.017)	155.421	8.847
aorsf-random	0.300 (0.032)	0.849 (0.027)	0.869	0.039
xgboost-aft	—	0.927 (0.010)	33.545	0.008
NCCTG Lung Cancer; death, n = 228, p = 9				
aorsf-random	0.076 (0.030)	0.686 (0.026)	0.304	0.015
ranger-extratrees	0.062 (0.028)	0.675 (0.033)	0.023	0.031
aorsf-fast	0.060 (0.043)	0.672 (0.026)	0.067	0.016
obliqueRSF-net	0.058 (0.026)	0.676 (0.029)	106.084	1.468
aorsf-cph	0.058 (0.041)	0.670 (0.024)	0.150	0.016
cif-standard	0.055 (0.032)	0.670 (0.030)	0.130	0.252
cif-extension	0.051 (0.032)	0.664 (0.029)	3.799	1.502
aorsf-net	0.047 (0.040)	0.668 (0.026)	16.041	0.016
glmnet-cox	0.033 (0.031)	0.638 (0.059)	0.081	0.002
rsf-standard	0.023 (0.039)	0.642 (0.025)	0.087	0.037
cif-rotate	0.017 (0.041)	0.632 (0.032)	4.687	1.897
xgboost-cox	0.011 (0.023)	0.648 (0.032)	0.721	0.002

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	-0.051 (0.163)	0.512 (0.082)	10.516	0.209
xgboost-aft	—	0.637 (0.034)	6.271	0.005
NKI 70 gene signature; death or metastasis, n = 144, p = 77				
aorsf-net	0.142 (0.056)	0.804 (0.056)	10.798	0.014
aorsf-cph	0.124 (0.049)	0.802 (0.051)	0.078	0.013
aorsf-fast	0.121 (0.052)	0.802 (0.054)	0.052	0.014
cif-rotate	0.118 (0.059)	0.787 (0.049)	27.382	3.379
obliqueRSF-net	0.106 (0.051)	0.792 (0.061)	79.059	0.558
cif-extension	0.098 (0.055)	0.799 (0.061)	8.595	3.554
cif-standard	0.088 (0.051)	0.781 (0.065)	0.145	0.149
rsf-standard	0.087 (0.048)	0.755 (0.050)	0.067	0.025
ranger-extratrees	0.064 (0.044)	0.774 (0.054)	0.024	0.030
aorsf-random	0.053 (0.045)	0.741 (0.060)	0.183	0.015
nn-cox	0.051 (0.102)	0.715 (0.100)	11.692	0.121
glmnet-cox	0.049 (0.064)	0.726 (0.090)	0.265	0.002
xgboost-cox	-0.028 (0.029)	0.566 (0.093)	0.118	0.002
xgboost-aft	—	0.770 (0.056)	4.896	0.005
Non-alcohol fatty liver disease; death, n = 17549, p = 24				
aorsf-cph	0.213 (0.008)	0.868 (0.005)	17.724	1.232
aorsf-fast	0.212 (0.009)	0.869 (0.005)	4.868	1.247
aorsf-net	0.210 (0.008)	0.864 (0.006)	471.404	1.283
obliqueRSF-net	0.209 (0.008)	0.868 (0.006)	1428.698	1042.701
rsf-standard	0.207 (0.009)	0.860 (0.005)	10.516	1.205
glmnet-cox	0.207 (0.011)	0.860 (0.005)	1.345	0.012
cif-standard	0.205 (0.007)	0.863 (0.006)	67.597	624.624
cif-rotate	0.190 (0.008)	0.865 (0.005)	263.660	62.919
ranger-extratrees	0.181 (0.007)	0.860 (0.005)	39.632	80.768
cif-extension	0.166 (0.003)	0.866 (0.006)	125.288	53.237
aorsf-random	0.140 (0.006)	0.838 (0.007)	8.915	1.339
xgboost-cox	0.020 (0.015)	0.876 (0.005)	8.907	0.017
nn-cox	-0.002 (0.009)	0.565 (0.092)	20.937	106.131
xgboost-aft	—	0.875 (0.005)	27.908	0.015
Primary biliary cholangitis; death, n = 276, p = 19				
aorsf-fast	0.430 (0.032)	0.908 (0.021)	0.082	0.019
aorsf-cph	0.418 (0.034)	0.906 (0.021)	0.162	0.018
aorsf-net	0.413 (0.035)	0.905 (0.021)	14.047	0.019
cif-rotate	0.405 (0.040)	0.899 (0.022)	10.102	1.933
rsf-standard	0.392 (0.034)	0.895 (0.023)	0.113	0.038
obliqueRSF-net	0.369 (0.032)	0.907 (0.022)	111.656	1.787
aorsf-random	0.354 (0.031)	0.893 (0.020)	0.308	0.020
cif-standard	0.352 (0.034)	0.904 (0.025)	0.206	0.363
cif-extension	0.348 (0.033)	0.901 (0.023)	5.775	2.199
glmnet-cox	0.342 (0.044)	0.886 (0.028)	0.116	0.002
ranger-extratrees	0.277 (0.027)	0.894 (0.027)	0.029	0.038
xgboost-cox	0.256 (0.103)	0.881 (0.027)	4.960	0.003

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	-0.017 (0.018)	0.563 (0.124)	11.751	0.247
xgboost-aft	—	0.883 (0.024)	9.644	0.006
Rotterdam tumor bank; death, n = 2982, p = 11				
aorsf-net	0.166 (0.012)	0.762 (0.009)	147.998	0.185
obliqueRSF-net	0.163 (0.010)	0.761 (0.009)	439.210	37.683
aorsf-cph	0.163 (0.012)	0.759 (0.009)	2.520	0.201
aorsf-fast	0.161 (0.012)	0.757 (0.009)	0.845	0.198
cif-standard	0.159 (0.010)	0.759 (0.009)	4.629	22.523
rsf-standard	0.159 (0.014)	0.756 (0.009)	3.066	0.967
aorsf-random	0.153 (0.010)	0.752 (0.010)	2.840	0.180
cif-rotate	0.147 (0.011)	0.751 (0.011)	34.328	8.629
ranger-extratrees	0.139 (0.006)	0.749 (0.009)	3.498	2.152
xgboost-cox	0.130 (0.014)	0.753 (0.010)	4.068	0.004
cif-extension	0.129 (0.004)	0.751 (0.008)	22.740	8.982
glmnet-cox	0.118 (0.008)	0.731 (0.009)	0.278	0.004
nn-cox	-0.009 (0.042)	0.531 (0.050)	16.635	8.901
xgboost-aft	—	0.761 (0.009)	14.813	0.006
Rotterdam tumor bank; recurrence, n = 2982, p = 11				
obliqueRSF-net	0.148 (0.010)	0.737 (0.009)	520.813	39.774
aorsf-net	0.146 (0.011)	0.735 (0.009)	160.272	0.191
aorsf-cph	0.145 (0.012)	0.734 (0.008)	2.678	0.205
cif-standard	0.144 (0.011)	0.734 (0.009)	4.807	23.027
aorsf-fast	0.143 (0.011)	0.733 (0.009)	0.854	0.206
aorsf-random	0.141 (0.010)	0.730 (0.008)	3.192	0.185
rsf-standard	0.139 (0.012)	0.731 (0.008)	3.003	1.000
ranger-extratrees	0.135 (0.007)	0.734 (0.009)	3.564	2.555
cif-rotate	0.129 (0.010)	0.725 (0.009)	36.519	8.483
cif-extension	0.119 (0.006)	0.731 (0.008)	23.055	8.825
glmnet-cox	0.117 (0.008)	0.727 (0.008)	0.258	0.004
xgboost-cox	0.113 (0.008)	0.729 (0.009)	3.628	0.004
nn-cox	-0.007 (0.027)	0.511 (0.045)	17.722	7.825
xgboost-aft	—	0.735 (0.009)	14.317	0.006
Serum free light chain; death, n = 7874, p = 10				
aorsf-fast	0.250 (0.014)	0.825 (0.007)	2.023	0.587
aorsf-cph	0.250 (0.013)	0.825 (0.008)	6.401	0.580
aorsf-net	0.250 (0.012)	0.823 (0.008)	278.411	0.563
glmnet-cox	0.248 (0.012)	0.820 (0.007)	0.539	0.006
obliqueRSF-net	0.247 (0.011)	0.821 (0.008)	1113.284	151.041
ranger-extratrees	0.243 (0.009)	0.820 (0.007)	12.743	12.085
cif-standard	0.243 (0.011)	0.818 (0.008)	19.565	120.023
rsf-standard	0.243 (0.013)	0.815 (0.008)	5.399	0.585
aorsf-random	0.231 (0.012)	0.816 (0.008)	6.681	0.568
cif-rotate	0.228 (0.009)	0.819 (0.007)	65.379	21.402
cif-extension	0.201 (0.005)	0.820 (0.008)	41.381	20.976
xgboost-cox	0.094 (0.038)	0.824 (0.007)	5.927	0.008

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	0.002 (0.005)	0.616 (0.122)	26.319	26.895
xgboost-aft	—	0.823 (0.008)	18.504	0.009
SPRINT; CVD death, n = 9361, p = 174				
glmnet-cox	0.071 (0.011)	0.795 (0.011)	13.850	0.011
aorsf-net	0.070 (0.007)	0.796 (0.011)	337.860	0.660
aorsf-fast	0.069 (0.006)	0.797 (0.011)	2.581	0.638
aorsf-cph	0.069 (0.006)	0.797 (0.011)	9.028	0.596
obliqueRSF-net	0.068 (0.005)	0.798 (0.012)	1017.350	425.134
rsf-standard	0.065 (0.007)	0.788 (0.014)	3.894	1.314
cif-standard	0.061 (0.003)	0.798 (0.011)	50.694	181.616
cif-rotate	0.060 (0.005)	0.791 (0.012)	930.893	113.261
ranger-extratrees	0.054 (0.003)	0.791 (0.012)	7.314	7.943
nn-cox	0.039 (0.018)	0.764 (0.027)	20.414	24.555
cif-extension	0.034 (0.002)	0.789 (0.011)	122.779	33.342
aorsf-random	0.026 (0.002)	0.747 (0.016)	5.887	0.742
xgboost-cox	0.006 (0.017)	0.799 (0.011)	6.888	0.013
xgboost-aft	—	0.796 (0.012)	20.126	0.012
SPRINT; death, n = 9361, p = 174				
glmnet-cox	0.123 (0.012)	0.771 (0.009)	5.422	0.010
aorsf-cph	0.117 (0.008)	0.770 (0.008)	12.876	1.527
aorsf-fast	0.116 (0.008)	0.770 (0.008)	3.583	1.514
aorsf-net	0.113 (0.009)	0.769 (0.009)	590.419	0.934
obliqueRSF-net	0.112 (0.007)	0.767 (0.008)	2630.588	231.667
rsf-standard	0.110 (0.008)	0.763 (0.009)	6.407	0.663
cif-standard	0.106 (0.006)	0.764 (0.008)	49.842	190.804
nn-cox	0.097 (0.010)	0.757 (0.009)	34.063	33.278
ranger-extratrees	0.096 (0.005)	0.756 (0.009)	11.323	9.400
cif-rotate	0.090 (0.007)	0.745 (0.009)	1109.925	112.893
cif-extension	0.055 (0.002)	0.747 (0.009)	137.470	34.230
aorsf-random	0.052 (0.003)	0.720 (0.010)	9.559	1.034
xgboost-cox	0.030 (0.023)	0.772 (0.008)	9.057	0.014
xgboost-aft	—	0.772 (0.007)	23.335	0.013
Systolic Heart Failure; death, n = 2231, p = 41				
obliqueRSF-net	0.114 (0.012)	0.747 (0.012)	381.891	25.692
glmnet-cox	0.113 (0.013)	0.745 (0.012)	0.276	0.004
cif-rotate	0.113 (0.013)	0.741 (0.011)	71.714	10.523
aorsf-net	0.112 (0.013)	0.743 (0.012)	118.505	0.158
aorsf-cph	0.112 (0.013)	0.744 (0.012)	1.895	0.147
aorsf-fast	0.110 (0.015)	0.743 (0.011)	0.586	0.146
cif-standard	0.110 (0.011)	0.744 (0.011)	4.079	16.213
rsf-standard	0.105 (0.011)	0.735 (0.011)	2.510	0.252
cif-extension	0.094 (0.006)	0.744 (0.012)	29.918	10.109
ranger-extratrees	0.091 (0.008)	0.738 (0.012)	3.266	1.328
xgboost-cox	0.090 (0.009)	0.744 (0.010)	4.229	0.004
aorsf-random	0.080 (0.006)	0.731 (0.013)	2.448	0.151

A.2: Index of prediction accuracy, time-dependent concordance statistic, and computational time required to fit and compute predictions for several learning algorithms across 31 risk prediction tasks. *(continued)*

	Scaled Brier	C-Statistic	Model fitting	Risk prediction
nn-cox	0.074 (0.028)	0.705 (0.036)	18.945	5.181
xgboost-aft	—	0.741 (0.009)	13.087	0.006
VA lung cancer trial; death, n = 137, p = 8				
aorsf-net	0.201 (0.050)	0.797 (0.035)	9.602	0.011
aorsf-fast	0.200 (0.050)	0.795 (0.034)	0.082	0.014
cif-rotate	0.198 (0.065)	0.789 (0.036)	4.476	1.284
aorsf-cph	0.198 (0.052)	0.794 (0.035)	0.105	0.011
rsf-standard	0.176 (0.048)	0.787 (0.037)	0.078	0.025
cif-extension	0.174 (0.048)	0.795 (0.034)	3.676	1.195
glmnet-cox	0.160 (0.036)	0.788 (0.037)	0.087	0.002
aorsf-random	0.154 (0.044)	0.780 (0.037)	0.213	0.012
cif-standard	0.128 (0.040)	0.770 (0.037)	0.105	0.120
obliqueRSF-net	0.126 (0.034)	0.796 (0.029)	62.935	0.664
ranger-extratrees	0.092 (0.033)	0.778 (0.038)	0.020	0.026
xgboost-cox	0.067 (0.078)	0.750 (0.046)	1.408	0.002
xgboost-aft	—	0.754 (0.047)	5.530	0.005
nn-cox	-0.023 (0.033)	0.517 (0.114)	11.344	0.133

A.3: Discrimination of relevant versus irrelevant variables for several techniques to estimate variable importance.

		accelerated oblique RSF			xgboost		RSF
Max correlation	No. observations	Negation	ANOVA	Permutation	SHAP	Gain	Permutation
Overall	Overall	76.0	74.0	73.2	69.7	64.7	67.8
Interactions							
Overall	Overall	57.9	57.8	58.1	54.7	49.4	57.1
30	500	54.3	54.5	54.7	48.2	42.6	54.8
30	1,000	57.0	56.6	57.9	53.5	48.7	56.5
30	2,500	61.9	59.2	64.2	61.9	61.2	60.3
15	500	53.4	53.6	53.1	47.0	40.9	54.5
15	1,000	56.5	55.7	57.1	52.2	45.9	55.5
15	2,500	61.3	59.0	62.6	61.1	59.0	60.8
0	500	52.6	54.8	53.2	44.6	40.8	53.6
0	1,000	56.9	58.7	55.5	53.0	43.1	55.6
0	2,500	67.6	68.1	64.6	71.1	62.3	62.4
Non-linear effects							
Overall	Overall	71.8	69.4	67.9	65.9	60.1	61.7
30	500	58.6	58.6	57.4	53.1	48.6	55.4
30	1,000	61.2	59.6	58.6	57.4	52.0	55.9
30	2,500	62.1	60.1	61.0	60.1	56.4	58.1
15	500	63.6	61.5	60.8	54.6	49.0	57.7
15	1,000	67.6	64.9	64.8	62.4	55.8	59.1
15	2,500	70.7	67.3	68.8	66.7	62.1	62.6
0	500	75.6	72.5	69.3	60.0	56.2	61.2
0	1,000	88.4	84.0	78.3	81.7	69.1	67.2
0	2,500	98.3	96.2	91.9	97.5	91.4	78.5
Combination effects							
Overall	Overall	78.4	76.0	74.9	70.8	65.3	68.3
30	500	64.8	63.8	62.7	55.6	49.8	59.3
30	1,000	67.5	65.7	65.0	61.4	55.7	61.5
30	2,500	69.8	67.0	68.5	65.7	62.4	64.3
15	500	70.4	68.0	66.6	59.2	53.0	62.3
15	1,000	75.0	71.2	71.7	66.6	59.7	64.9
15	2,500	78.9	74.8	77.1	72.9	69.0	69.9
0	500	84.0	81.4	76.5	66.5	61.7	67.4
0	1,000	95.4	92.4	87.8	89.4	78.6	76.3

A.3: Discrimination of relevant versus irrelevant variables for several techniques to estimate variable importance. *(continued)*

Max correlation	No. observations	Negation	ANOVA	Permutation	SHAP	Gain	Permutation
0	2,500	99.8	99.3	97.9	99.6	97.8	89.1
Main effects							
Overall	Overall	91.1	89.0	88.8	85.1	82.7	83.2
30	500	79.3	77.5	75.3	70.2	66.5	71.2
30	1,000	83.7	80.8	80.6	77.1	74.0	74.8
30	2,500	86.7	83.7	85.2	81.8	80.5	79.3
15	500	86.4	83.3	81.9	75.6	71.3	75.4
15	1,000	91.3	88.2	88.7	84.9	81.6	81.0
15	2,500	94.6	91.8	93.8	90.3	89.0	86.8
0	500	97.7	96.2	94.1	86.2	83.6	85.7
0	1,000	100.0	99.7	99.4	99.4	98.0	95.1
0	2,500	100.0	100.0	100.0	100.0	100.0	99.8

References

Karel GM Moons, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward. Risk prediction models: II. external validation, model updating, and impact assessment. *Heart*, 98(9):691–698, 2012a.

Karel GM Moons, Andre Pascal Kengne, Mark Woodward, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Diederick E Grobbee. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98(9):683–690, 2012b.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

Hong Wang and Gang Li. A selective review on random survival forests for high dimensional data. *Quantitative bio-science*, 36(2):85, 2017.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064, 2010.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Consistency of survival tree and forest models: splitting bias and correction. *arXiv preprint arXiv:1707.09631*, 2017.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Bjoern H Menze, B Michael Kelm, Daniel N Splitthoff, Ullrich Koethe, and Fred A Hamprecht. On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 453–469. Springer, 2011.

Byron C Jaeger, D Leann Long, Dustin M Long, Mario Sims, Jeff M Szychowski, Yuan-I Min, Leslie A Mcclure, George Howard, and Noah Simon. Oblique random survival forests. *The Annals of Applied Statistics*, 13(3):1847–1883, 2019.

David Heath, Simon Kasif, and Steven Salzberg. Induction of oblique decision trees. In *IJCAI*, volume 1993, pages 1002–1007. Citeseer, 1993.

Sreerama K Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994.

Hemant Ishwaran and Min Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38(4):558–582, 2019.

Le Zhang and Ponnuthurai N Suganthan. Oblique decision tree ensemble via multisurface proximal support vector machine. *IEEE transactions on cybernetics*, 45(10):2165–2176, 2014.

Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.

Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

Nitesh Poona, Adriaan Van Niekerk, and Riyad Ismail. Investigating the utility of oblique tree-based ensembles for the classification of hyperspectral data. *Sensors*, 16(11):1918, 2016.

Xueheng Qiu, Le Zhang, Ponnuthurai Nagaratnam Suganthan, and Gehan AJ Amaratunga. Oblique random forest ensemble via least square estimation for time series forecasting. *Information Sciences*, 420:249–262, 2017.

Tyler M Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L Patsolic, Benjamin Falk, Carey E Priebe, Jason Yim, Randal Burns, Mauro Maggioni, et al. Sparse projection oblique randomer forests. *Journal of machine learning research*, 21(104), 2020.

Rakesh Katuwal, Ponnuthurai Nagaratnam Suganthan, and Le Zhang. Heterogeneous oblique random forest. *Pattern Recognition*, 99:107078, 2020.

Ruoqing Zhu. *Tree-based Methods for Survival Analysis and High-dimensional Data*. PhD thesis, The University of North Carolina at Chapel Hill, 2013.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Lifeng Zhou, Hong Wang, and Qingsong Xu. Random rotation survival forest for high dimensional censored data. *SpringerPlus*, 5(1):1–10, 2016.

Hong Wang and Lifeng Zhou. Random survival forest with space extensions for censored data. *Artificial intelligence in medicine*, 79:52–61, 2017.

H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2019. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.8.0, available at `https://cran.r-project.org/package=randomForestSRC`.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(1):1–17, 2017. doi:10.18637/jss.v077.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v077i01`.

Torsten Hothorn, Kurt Hornik, Carolin Strobl, and Achim Zeileis. Party: a laboratory for recursive partytioning, 2010.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

Terry Therneau. Survival package source code documentation, April 2022a. URL `https://github.com/therneau/survival/blob/5440691d44abea537b08aeb60153a31654d66a9b/noweb`. original-date: 2016-04-28.

Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi:10.1001/jama.1982.03320430047030. URL `https://doi.org/10.1001/jama.1982.03320430047030`.

William Michael Landau. The targets r package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, 6(57):2959, 2021. URL `https://doi.org/10.21105/joss.02959`.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999. URL `https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18%3C2529::AID-SIM274%3E3.0.CO;2-5`.

Michael W Kattan and Thomas A Gerds. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2(1):1–7, 2018.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, 32(30):5381–5397, 2013.

ARIC Investigators. The atherosclerosis risk in communit (aric) study: design and objectives. *American journal of epidemiology*, 129(4):687–702, 1989.

Diane E Bild, David A Bluemke, Gregory L Burke, Robert Detrano, Ana V Diez Roux, Aaron R Folsom, Philip Greenland, David R JacobsJr, Richard Kronmal, Kiang Liu, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881, 2002.

G Michael Felker, Kevin J Anstrom, Kirkwood F Adams, Justin A Ezekowitz, Mona Fiuzat, Nancy Houston-Miller, James L Januzzi, Daniel B Mark, Ileana L Piña, Gayle Passmore, et al. Effect of natriuretic peptide–guided therapy on hospitalization or cardiovascular mortality in high-risk patients with heart failure and reduced ejection fraction: a randomized clinical trial. *Jama*, 318(8):713–720, 2017.

SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *The Journal of Machine Learning Research*, 18 (1):2653–2688, 2017.

Max Kuhn and Hadley Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL `https://www.tidymodels.org`.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2022. URL `https://mc-stan.org/rstanarm/`. R package version 2.21.3.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Glenn Heller. A measure of explained risk in the proportional hazards model. *Biostatistics*, 13(2):315–325, 2012.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

Charles G Moertel, Thomas R Fleming, John S Macdonald, Daniel G Haller, John A Laurie, Catherine M Tangen, James S Ungerleider, William A Emerson, Douglass C Tormey, John H Glick, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of internal medicine*, 122(5):321–326, 1995.

Terry M Therneau. *A Package for Survival Analysis in R*, 2022b. URL `https://CRAN.R-project.org/package=survival`. R package version 3.3-1.

Terry M Therneau and Patricia M Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.

Byron C Jaeger. *aorsf: Accelerated Oblique Random Survival Forests*, 2022. URL `https://github.com/bcjaeger/aorsf`. R package version 1.0.0.

Emil Hvitfeldt and Hannah Frick. *censored: 'parsnip' Engines for Survival Models*. URL `https://github.com/tidymodels/censored`. R package version 0.1.0.9000.

M Schumacher. Rauschecker for the german breast cancer study group, randomized 2 by 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12:2086–2093, 1994.

Torsten Hothorn. *TH.data: TH's Data Archive*, 2022. URL `https://CRAN.R-project.org/package=TH.data`. R package version 1.1-1.

Eileen Hsich, Eiran Z Gorodeski, Eugene H Blackstone, Hemant Ishwaran, and Michael S Lauer. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes*, 4(1):39–45, 2011.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton III, et al. Use of nonclonal serum

immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, pages 517–523. Elsevier, 2012.

Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Dirk R Larson, Matthew F Plevak, Janice R Offord, Angela Dispenzieri, Jerry A Katzmann, and L Joseph Melton III. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 354(13):1362–1369, 2006.

Alina M Allen, Terry M Therneau, Joseph J Larson, Alexandra Coward, Virend K Somers, and Patrick S Kamath. Nonalcoholic fatty liver disease incidence and impact on metabolic burden and death: a 20 year-community study. *Hepatology*, 67(5):1726–1736, 2018.

Patrick Royston and Douglas G Altman. External validation of a cox prognostic model: principles and methods. *BMC medical research methodology*, 13(1):1–15, 2013.

David W Hosmer and Stanley Lemeshow. *Applied survival analysis: regression modelling of time to event data*. Wiley, 2002.

Raphael Sonabend, Franz J Király, Andreas Bender, Bernd Bischl, and Michel Lang. mlr3proba: An r package for machine learning in survival analysis. *Bioinformatics*, 02 2021. ISSN 1367-4803. doi:10.1093/bioinformatics/btab039.

Christine Desmedt, Angelo Di Leo, Evandro de Azambuja, Denis Larsimont, Benjamin Haibe-Kains, Jean Selleslags, Suzette Delaloge, Caroline Duhem, Jean-Pierre Kains, Birgit Carly, et al. Multifactorial approach to predicting resistance to anthracyclines. *Journal of Clinical Oncology*, 29(12):1578–1586, 2011.

Christos Hatzis, Lajos Pusztai, Vicente Valero, Daniel J Booser, Laura Esserman, Ana Lluch, Tatiana Vidaurre, Frankie Holmes, Eduardo Souchon, Hongkun Wang, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Jama*, 305(18):1873–1881, 2011.

Nils Ternès, Federico Rotolo, Georg Heinze, and Stefan Michiels. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal*, 59(4):685–701, 2017.

Nils Ternes, Federico Rotolo, and Stefan Michiels. *biospear: Biomarker Selection in Penalized Regression Models*, 2018. URL `https://CRAN.R-project.org/package=biospear`. R package version 1.0.2.

Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

Giuseppe Casalicchio, Jakob Bossek, Michel Lang, Dominik Kirchhoff, Pascal Kerschke, Benjamin Hofner, Heidi Seibold, Joaquin Vanschoren, and Bernd Bischl. OpenML: An R package to connect to the machine learning platform OpenML. *Computational Statistics*, pages 1–15, 2017. doi:10.1007/s00180-017-0742-2. URL `http://dx.doi.org/10.1007/s00180-017-0742-2`.

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

Charles Lawrence Loprinzi, John A Laurie, H Sam Wieand, James E Krook, Paul J Novotny, John W Kugler, Joan Bartel, Marlys Law, Marilyn Bateman, and Nancy E Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–607, 1994.

Melania Pintilie. *Competing risks: a practical perspective*. John Wiley & Sons, 2006.

Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, Matthew F Plevak, and L Joseph Melton III. A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England Journal of Medicine*, 346(8):564–569, 2002.